Dynamics of Trending Topics

Smart Service Systems using Trending Topics

UNIVERSITY OF TASMANIA

Soyeon Caren Han

BComp(Hons, First Class)
University of Tasmania

A dissertation submitted to the Faculty of Engineering and ICT, University of Tasmania in fulfillment of the requirements for the degree of
Doctor of Philosophy

April 2016

Declaration

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Soyeon Caren Han
April 2016

Authority

Soyeon Caren Han
April 2016

Ethics

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University

Soyeon Caren Han
April 2016

I would like to dedicate this thesis to my loving husband, David Chung …

Acknowledgements

Firstly, I would like to express my grateful appreciation to my supervisor, Associate Prof. Byeong Ho Kang, for the constant support of my PhD study and related research, for his patience, motivation, and immense knowledge. Without his guidance and efforts in all the time of research, I could not finish this research and thesis. I could not have imagined having a better advisor and mentor for my PhD study.

Besides my supervisor, I would like to thank to my co-supervisor, Associate Prof. Paul Turner for his insightful comments and encouragement. My grateful thanks also goes to Dr. Yangsok Kim, Prof. Paul Compton and Prof. Hiroshi Motoda who provided me an opportunity to receive their precious feedbacks of my work, which guide me to see more insight of this research.

I thank my fellow lab mates in Smart Service Systems group and my friends. In particular, I would like to thank all my special friends, Leo Chen, Grace Jung, and Sze Yen.

Last but not the least, I would like to thank my family: my husband, David Chung and my parents for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

What is the definition of trending topic? Trending topic is the topic that a great amount of people are interested in. While trending topics could be detected from the offline survey or newspaper in the past, those topics can be easily detected from online these days. The rise of new types of social networking services, such as Facebook or Twitter, has caused the accumulation of unprecedented amount of web social data. This large amount of social data attracts many researchers' and companies' attention since it enables to collect and monitor the social interests with reducing the time and cost for the quantitative survey. Almost all web and social networking services analyze their user created social data and detect the most popular terms that are discussed and searched by their users. The popular terms are detected and published in the list, called 'Trending Topic' list. Awareness and utilization of trending topics plays a crucial role in various fields, including marketing, politics, and economics.

This dissertation focuses on the following three studies. The first study is to analyse the nature of trending topics and reveal the important aspects. The second study is to identify the relevance of trending topic to a target object, such as individuals or organisations. It discovers how a trending topic affects a target group or individual. The impact strength of a trending topic is closely related to the strength of the relationship between the user and the issue. Therefore, this study will propose a method to identify the relevance between users and trending topics. The last study is to develop the model that predicts the trends of trending topics in the future. This study focuses on the formation and fading of trending topics and the trend during this lifecycle. Moreover, the dissertation also propose a model that represents the diffusion drifting movement of trending topic among different online communities.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

This dissertation focused on investigating the dynamics of trending topics, and proposing new type of smart services framework using trending topics. The opening chapter will contain the main motivations for this process in chapter 1.1, and the summary of dissertation outlines in chapter 1.2

1.1 Motivation

The rises of new types of social networking services, such as Facebook and Twitter, have caused a human communication paradigm shift by increasing personal information sharing. The phenomenon, called "Social data revolution", has resulted in the accumulation of unprecedented amount of social data. This large amount of social data, which is made by users themselves, is like a vein of gold in 21st century so it attracts many researchers' and companies' attention. Why social data become an important asset for both companies and government? This is because social data represents what people are currently are doing and thinking (are interested in). By analyzing those social data, it is possible to collect and monitor the social interests and behaviour with reducing the time and cost for the quantitative survey.

Not surprisingly, many large Internet-based companies already establish the social data analytics research lab and provide the analytics result to the public through various analytics services. One of the most popular analytics services is the "Trending Topic service", which presents the online trending topics that most people are currently discussed/searched/read. This service is provided by different large internet-based companies, including Search engine, Social Media, and Internet news. For example, Twitter provides the service 'Twitter Trending Topics ', which displays the list of top 10 fastest rising discussed terms in the Twitter. These 'top 10 discussed terms' are updated in the real time.

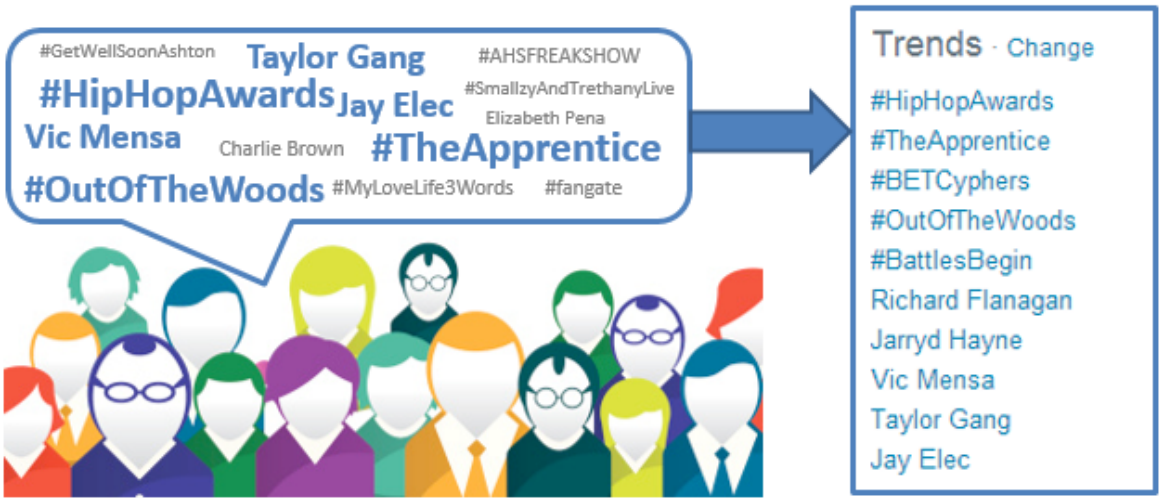

Fig. 1.1 Trending Topics in Social Media

Trending Topics are estimated to reflect the real-world issues from the people's point of view. For example, Kwak et al. (Kwak, 2010) indicated that over 85% of trending topics in Twitter are related to breaking news headlines, and the related tweets of each trending topic provides more detailed information of people's opinions. Being able to recognize and utilize the trending topics, people are currently most interested in online web communities, may lead to opportunities for analyzing the market share in almost every industry and research fields, including marketing, politics, and economics. Most Internet users are not aware of the latent ability of those services but noticing trending topic enables a lot of opportunities for people and organisations. If the user, an individual or an organization, can define the related trending topic and predict the future trends of them, they can be prepared for the future impact of that topic.

Through this dissertation, the following three studies are included. The first study is to characterise the trending topics by tracking the trending topics in different countries and services. We also propose a model to disambiguate the meaning of trending topics. The second study is to identify the relevance of trending topic to a target object, such as individuals or organisations. It discovers how a trending topic affects a target group or individual. The impact strength of a trending topic is closely related to the strength of the relationship between the user and the issue. Therefore, this study will propose a method to identify the relevance between users and trending topics. The final study is to develop the model that predicts the trends of trending topics in the future. This study focuses on the formation and fading of trending topics and the trend during this lifecycle. Moreover, i also propose a model that represents the diffusion drifting movement of online trending topic.

The motivations for each chapter can be found from its introduction chapter.

1.2 Thesis Overview

The dissertation aims to investigate analysing the nature of trending topics, to propose the new framework for identifying the relevance of trending topics to target domain, and to develop the framework for forecasting lifetime and diffusion trends of trending topics among different online communities, including web services community, and country-based online community.

The following list includes three main research questions:

1. How to analyse and reveal the nature of trending topics?
   (a) How to characterise the trending topics
   (b) How to disambiguate the exact meaning of the trending topic from trending topic data analytic service?
2. How to identify the relevance of a trending topic to a target, such as individuals or organisations?
   (a) How to monitor the trending trending topic?
   (b) How to disambiguate the exact meaning of the trending topic from trending topic data analytic service?
   (c) How to monitor/manage the resources that represent the information and activity of a target?
   (d) How to calculate the relevance of a trending topic to a target domain?
   (e) How to identify the most related document/folder in the target domain
3. How to forecast the future trends of trending topic?
   (a) How to forecast the patterns of trending topic
   (b) How to predict the topic drifting of trending topics among different trends services
   (c) How to predict the trending topic diffusion among different countries

Figure 1.2 shows the dissertation overview, and summarises the structure and flow of the dissertation.

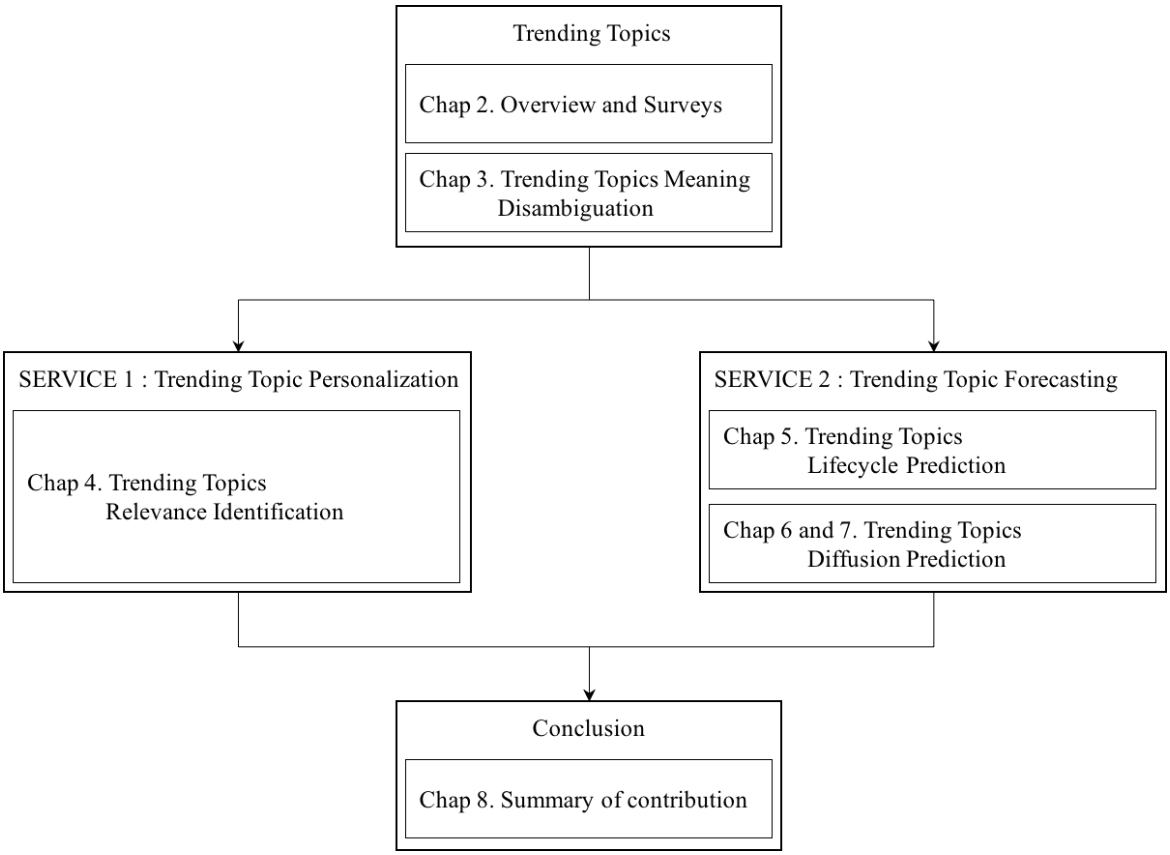

Fig. 1.2 Thesis Overview

The following list contains the detailed explanation of the contents of this dissertation.

- Chapter 1 contains the motivation of the research and a dissertation overview.
- Chapter 2 reviews previous research in prediction and recommendation approach using web social data, prediction and detection of trending topics using web social data, and the overview of different types of trending topics analytics services.

Chapter 3 is for analysing the trending topics and revealing the important aspect of trending topics for the further research.

- Chapter 3 contains the evaluation study in order to identify the best approach for trending topic meaning disambiguation by comparing four different information retrieval approaches, and it was evaluated by 20 postgraduate students. The best trending topic meaning disambiguation approach revealed in Chapter3 will be used for identifying the relevance of trending topic to a target domain in Chapter4, and also for predicting the future trend of trending topics lifecycle and diffusion in Chapters 5,6 and 7. Therefore, the Chapter 3 will be the starting point for the research in Chapters 4,5,6, and 7.

- Chapter 4 proposes new framework that identifies the relevance of trending topic to a certain target domain, including an individual or organisation. By identifying the personalised relevance, people or organisations are easy to get how much the certain trending topic is related to them.

Chapter 5, 6, 7 focuses on future trends of trending topics.

- Chapter 5 proposes new approach for trending topic lifecycle forecasting, and it proves that statistical analysis are possible to predict the future trends of trending topic lifecycle.

Trending topics show the popular topics among users in certain community (e.g. users in a certain web service or users in a certain country). Trending topics in one community can be different from others since the users in the community may discuss different topics from other communities. Surprisingly, i found that some trending topics are diffused among multiple communities In chapter 6 and 7, i focused on predicting the diffusion trends of trending topics among different online communities

- Chapter 6 investigates the characteristics of the trending topics in different types of trending topic analytics service, and proposes diffusion predicting model for forecasting how trending topics diffuses among three different online web services, including search engine, social media, and Internet news service.

- Chapter 7 introduces trending topic diffusion prediction model among online communities in different countries.

- Finally, Chapter 8 provides conclusions and future works, which summarise any contributions and recommend the future researches.

Chapter 2

Overview and survey

This chapter reviews some of the related researches that have been conducted for the prediction and recommendation approach using web social data, and prediction and detection of trending topics using web social data. It also provides the overview of different types of trending topics analytics services.

The detailed surveys and literature reviews can be found in each chpater.

2.1 Introduction

Social networking services (SNS) are the Webbased services that enable people to establish and maintain connections with their friends, family members, coworkers, etc. (Boyd & Ellison 2007). This service focuses on building online social communities that will allow people to share their interests and activities with others in real time (Liccardi et al. 2007). According to some researchers, SNS is the fastestgrowing type of social software. Due to the rapid development of SNS, it has drawn much attention of late and has become one of the main services of online networks (Richter & Koch 2008). Figure 2.1 indicates that the percentage of adults who use SNSs from 2005 to 2013. It is shown the dramatic increase.

There is no doubt that SNSs have drawn enormous interest in a short span of time. According to Takahashi (2010), social networking has become a main stream of webbased activity. It has changed people's lifestyles, and has influenced their handling of information (Kolbitsch & Maurer 2006). This has piqued the curiosity of many researchers. "What makes people very enthusiastic about SNS?" To find the answer to this question, some researchers have analysed several open and closed SNSs. In this chapter, the characteristics of SNSs are reviewed based on the research that has been done in this area. First of all, SNS enables people to keep in touch with others and to maintain their social capital. Until a few years ago, one of the main causes of the smaller social communities is moving to a new place on

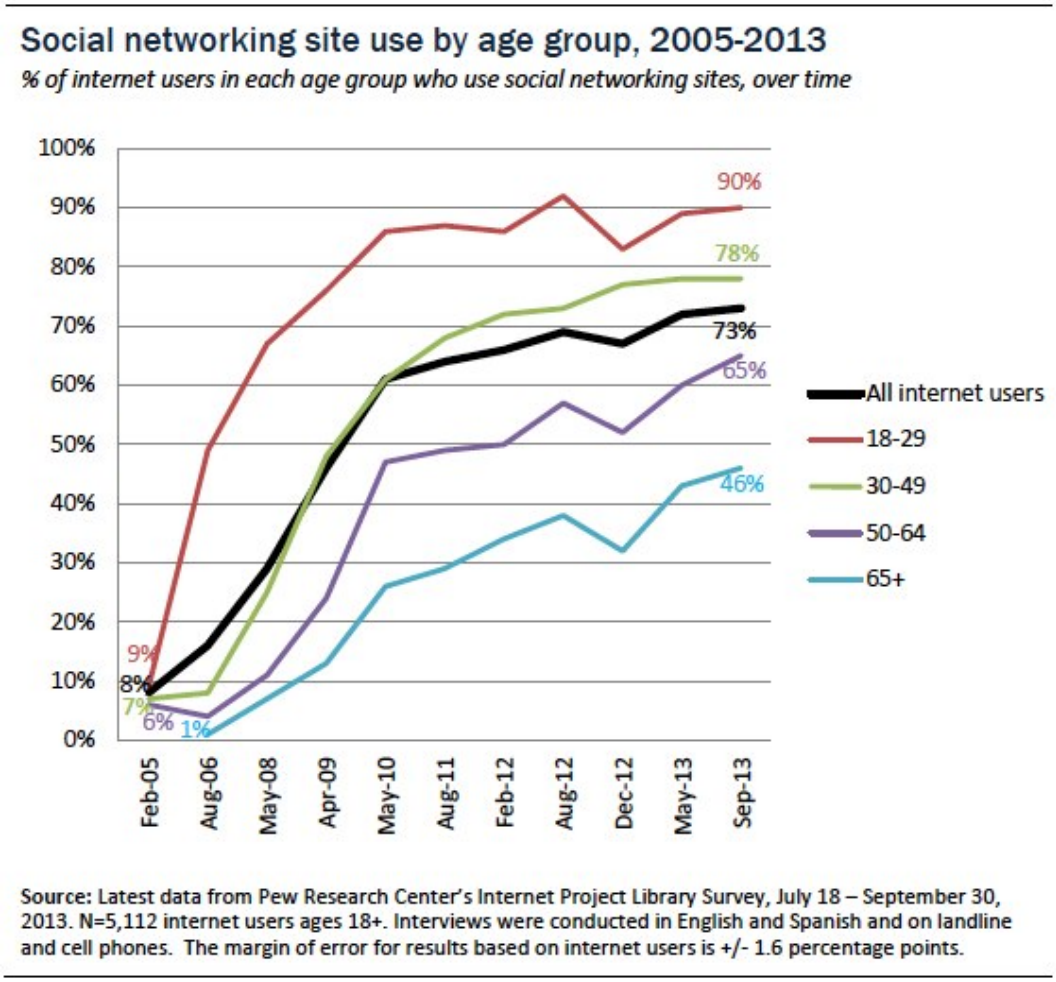

Fig. 2.1 The percentage of all adults SNS users

account of a new job (Putnam 2000). With the many kinds of online communication services that have been introduced, however, moving to a new place on account of a new job no longer makes social communities smaller as people can maintain longdistance relationships via Internet messenger services or SNSs. While Internet messenger services, however, require specific individual/group contacts, in many SNSs, the users do not need to contact target individuals. Unlike the common online communication services, SNS provides a new mechanism for users based on a profile page. When the users update their status in their profile page, including their contact numbers and other personal information, photos, and videos, the changes can be read only by the people who are already part of the users' respective extended social networks. In addition, this service offers a function that allows users to make comments on their friends' statuses (Boyd & Ellison 2007). Richter and Koch (2008) views SNS users as performers for particular audience in the stage.

2.2 Prediction Studies Using Social Data

In the last decades, there are a number of journals and conference papers, which are applied Twitter dataset in order to propose the new types of prediction and recommendation system in various domains, including election prediction, disease prediction, disaster prediction and detection, and trending topics prediction. The following section summarises different types of research papers aimed predicting election results in different location, from small city level to country level. The main goal of those papers is to identify whether it is possible to predict or detect people's opinions in certain politician by analysing their Twitter messages. The result of sentiment analysis is one of the good indicator in predicting election result. Several different machine learning techniques and sentimental analysis approaches are applied in the following research papers in order to use Twitter data in election prediction.

The most general approach is applying several types of different machine learning techniques in order to monitor, detect and forecast the political alignment among Twitter users (Conover, Goncalves, Ratkiewicz, Flammini and Menczer, 2011). In the research, researchers manually annotated training data, which are retrieved from 1000 different Twitter users. In order to learn the unique pattern (model) from the training dataset, support vector machine was applied. The learned model outputs the result, which identifies the twitter users in favour of left or right political ideologies based on the contents of user's tweets. The user generated metadata with latent semantic analysis is used for identifying the hidden sources of variation that are extermely related to the political alignment of users. In the same year, same researchers have used twitter Hashtags in order to classify the left right political spectrum. The research identifies the group of Twitter Hashtags, including two major HashTags (#p2

representing Progressive, #tcot representing Top Conservatives). Related retweeting and mentioning activities using those above hashtags were also monitored. The activities includes all communication and interaction among users. The evaluation results shows the users can be classified into the left and right political classes based on retweeting but mentioning activities. This is because retweeting activities are occurred based on the users' agreements and spreading through the online communities, while mentioning activities are the opposite.

Some researchers have aimed to analyse the contents, which are generated by the users in social media and identified people's opinion in the same style with tracking traiditional polls. The researchers measured and caculated the effects by analysing the sentimental class from public opinions. Analysing the sentimental class in twitter can be related to the polling dataset in the time series manner. Bermingham and Smeaton (2011) have developed on political sentimental analysis model in order to forecast the result of election. They applied supervised learning approach with volume related factor, share of volume (SOV), in analysing and classifying sentiment. The performance was evaluated by comparing the predicted and actual result. Several tool was implemented for monitoring the conversation with seed hashtags, and collecting the frequencies of predefined terms (Soler, Cuartero and Roblizo, 2012).

Various countries' election predictions are conducted. Choy, Cheong, Laik and Shung (2011) have researched on predicting the rate of voting in singapore election. They collected all those tweets posted during election campaign. From those collected dataset, some noise data, including ambiguous, irrelevant or repeating tweets, were filtered. Swedish election were also attempted to predict using twitter postings by Larsson and Moe (2010). The research found the unique type of users using extended twitter postings. The data was collected in before, during and after the election campaign in Sweden. Tumasjan, Sprenger, Sandner and Welpe (2011) have focused to perform sentiment analysis of tweets mentioning candidate or party name to evaluate the validity of Twitter data for predicting election results.

2.3 Trending Topics Detection Using Social Data

With the development of network, websites and online applications provide information that people may interested in. However, the data is too big to find out the real-time information or previous information, which requires researchers to extract the news from diverse webpages or datasets. As early as 1992, Andersen et al. proposed JASPER which applying template-driven method to extract news for solving significant business problems. The original intention of JASPER is to help the Reuters to analyse the financial news and reports which are provided by publicly-traded companies. Once the earnings and dividend reports are generated by

JASPER, the reporters only need check the necessary information, which helps the decision-makers make better decisions fast and accurately. Although Andersen et al. (1992) extracted information from text and table, they mentioned JASPER has a frame representation to check whether the new PR Newswire releases match the patterns and decides to assign a value to the slot. JASPER generates a new story from these information which is available for reporter to edit after the extracting and storing all available information. Andersen et al. (1992) try to evaluate the accuracy of extracted information for earnings and dividend releases provided by JASPER. They considered the measures of accuracy as completeness and correctness by testing 100 earnings and 50 dividend releases. Similarly, in order to extract relevant content, Laber et al. (2009) proposed NCE (News Content Extractor) to work. They indicated that their method works based on DOM tree representation of new web pages, which also applied in research of Reis et al. (2004) that extract information automatically from websites. Laber et al. (2009) assumed two hypotheses, which there is high measure of a node associated with the webpage and a positive real number, and which comments display after body of a news webpage. According to their observation, the measure of a news webpage achieves almost 90% by testing 324 news documents. However, in research of Reis et al. (2004), they rather study a specific type of tree called labelled ordered rooted tree. They presented a new algorithm for determining a new type of mapping called RTDM (Restricted Top-Down Mapping), which extracts information by page clustering, extraction pattern generation, data matching, and data labelling. They compared the extracted news by original HTML pages and by their approach from 35 sites. And the average 87.71% correctly results demonstrate the RTDM algorithm has highly effective for extracting new automatically.

Xia, Yu and Zhang (2009) also applied tree alignment algorithm for their research, proposed an automatic wrapper generation method. A heuristic method is employed for determining the most probable content block and the alignment algorithm detects repeating patterns on the union tree. Therefore, they compared their approach to RTDM which applied in research of Reis et al. (2004) by testing 9000 web pages including Blog, news, forums. The results show that the performance of proposed approach in blogs website is better than in news and forums websites. Although the results of their new tree alignment display out performances in Blog, news and forums website, the extracted information is complex and comprehensive. Ma & Wan (2010) provided an approach to classify only news comments from readers. Their approach aims to extract explicit and implicit opinion targets from news comments by based on Centring Theory. Ma & Wan (2010) extracted 'focused concepts and rank their importance by computing the semantic relatedness with sentences via Wikipedia'. The experiment demonstrates that the approach effective. However, their results are not obvious high accuracy. The information extraction not only are news websites, forums and

Blog, but also can be used in social media side such as Twitter. Medvet & Bartoli (2012) proposed an approach to detect popular topics, summarize these topics by the representation of their precise meanings, and evaluate sentiment polarity of each topic. Their approach employed with a given topic, which means they should collect the recent tweets related to that topic. After data collection, they identified the high quality of tweets and classified these tweets into three sentiment categories (positive, negative, or neutral). And then these representation tweets are summarized for each sentiment categories by Medvet & Bartoli (2012). They tested their approach to explain the precise meaningful qualitative evaluation of popular topics.

2.4 Trending Topics Analytic Service

As mentioned earlier, since the social media have been received much attention, tracking trends become the important issue in every field, such as knowledge acquisition or software engineering recently (Rech 2007). Most search engines and websites are providing the service that displays the trending topics. The method of social issue tracking can be classified into three main sections: searchbased, social networkingbased and newsbased social issue tracking. Firstly, the trends can be obtained by using the data on search phrases. Many search engines, such as Google or Yahoo, are providing the searchbased social issue tracking services. Secondly, SNSs, such as Twitter, enables user to track emerging trends. It is called by social networkingbased trend tracking method. Thirdly, many Internet news sites recommend the most popular stories to users based on clicking history. The most interesting issue in the trend tracking tools is that each tool uses different mechanisms and draws the different results. Even if they use the same tracking method, the result would be different. This is because each service collected different data.

2.4.1 Search Based Trending Topics Analytics

Most search engines provide the list of the most popular search terms of the moment. The lists are based on the data set of search results. The popular search keywords represent the topics that people are interested in. Most people tend to search information when they are really interested in some topic. Therefore, search based trends tacking tools are considered as the efficient way to understand the trends (Segev & Ahituv 2010).

The most notable example of search based social issue tracking tool is ‘Google Trends’ that is provided by Google as can be seen in the figure 2.2. Google is the most popular search engine that a lot of people use it to search information. By using the data from people,

Fig. 2.2 Hot Search Topics in Google Trends (Google Trends 2012)

Google offers the new service with a list of top 10 trending searchterms hourly. Even though Google Trends does not mention the exact methodology that they use, the service is regards as good indicator that enables user to understand the type of information that people are currently interested in (Tirado et al. 2011). Moreover, Kwak et al (2010) proves that the freshness of topics in Google Trends is much higher than other social issue tracking services.

2.4.2 SNS based Trending Topics Analytics

As mentioned earlier, social networking sites have been received so much attention. Since most SNS users tend to focus on sharing information and discovering the interesting issue from others (Java et al. 2007). Many websites did not pass this opportunity so that they provide what the most popular topic by using SNS messages. Recently, Twitter shows a list of the top 10 trending topics on the right side navigation bar as you can see the figure 2.3 . To track the most often discussed topics, Twitter collects all tweets, which contains phrases or words. As a result, the most often mentioned words are selected and posted in the list of the trending topics. Each Twitter user can access that service because it is available in each profile page. Unfortunately, like other websites do, Twitter does not release their exact algorithm that they use for trending topic service. (Kwak et al. 2010)

Since Twitter provides a lot of API (Application Programming Interface) for software engineers and researchers, there are several kinds of variance methods. For example, Sakaki, Okazaki & Matsuo (2010) have introduced Twitter based issue tracking to detect earthquake in Japan.

2.4.3 News based Social issue Analytics

Since internet news provides information in the quickest and direct way, many people prefer to use internet news services rather than newspaper. Many internet news sites, such as New York Times, news.com, or Google News, developed the topic recommended system.

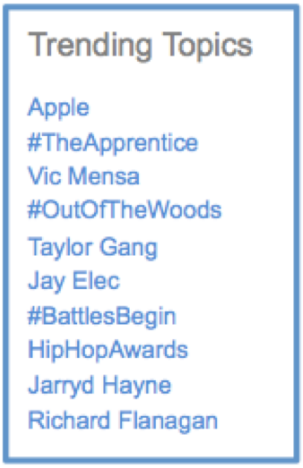

Fig. 2.3 Trending Topics in Twitter (Twitter 2012)

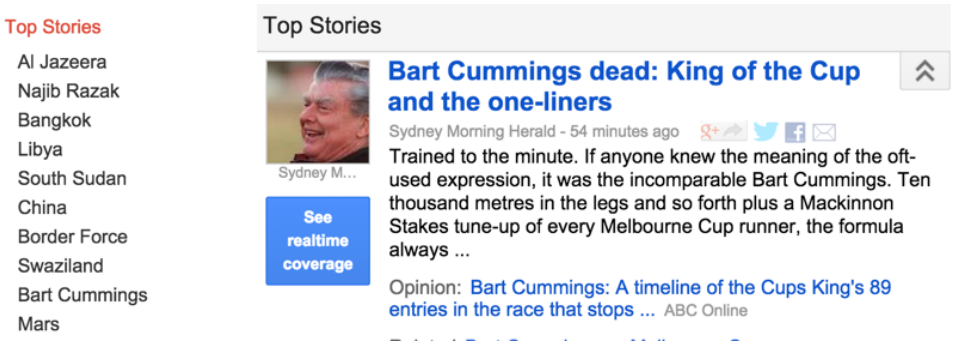

Fig. 2.4 Top Stories in Google News (Google News 2012)

Unlike other social issue tracking tools, Google News opens the news trends recommendation mechanism as can be seen in the figure 2.4. According to Liu, Dolan & Pedersen (2010), Google is developed a top stories visualisation system based on users clicking behaviour. First, they conducted a large amount of analysis of users' clicking behaviour. Then, it demonstrates the most popular news topics that reflect user's' current news interests.

2.5 Previous Researches in Trending Topics Analytic Service

Recently, social media are more and more popular for people to spread the news, events, and rumours, expressing interactions between individuals. People through following someone they interested in to know the important events, interesting news, their individual daily lives. In hence, the extracting information from news are not enough for current lifestyle. However,

these online social media services provide the current interested topics by their own servers or websites. As a rising microblogging service, Twitter has more than 41 million users as early as July 2009. The huge users mean that trending topics on Twitter can represent the opinions of more than 41 million users. As the most common online social media services, Twitter provides the real-time top 10 trending topics of different regions from cities to worldwide. Additional, Sina Weibo is a popular Chinese social media as same functions as Twitter, which not only presents the real-time top 10 trending keywords with their count of search and ranking, and rising trends in all kinds of tweets and each categories (containing current events, films and television, famous person, sports, and finance and economics), but also show the fastest rising hot words.

Despite these two social media, google trends and baidu provide more details for the search trends. Google trends displays real-time hot search in google like what Twitter and Sina Weibo presents. Moreover, it shows top 10 hot search for a specific day with each search count and top 10 hot search for integrated search or each categories. Once the hot search is focused on, google trends explore the trends of this topic among the time, the popularity for each region, and the related phrase for popularity and fastest rising. Baidu, as another search engine, displays more details and various categories of topics than google trends. Baidu not only provides the real-time hot search, but also displays a section called attention of 7 days, which presents the top 10 hot search for latest 7 days of integrated list. In contrast, Baidu divide each categories into several sub-categories, which lists 50 hot search for each subcategories (much more than top 10 hot search) with their own ranking and search count.

Kwak et al. (2010) presented the topological characteristics of Twitter. The trending topics are categorized by rank of users by their number of followers, by PageRank and by number of retweets. In their study, the results of using by number of followers and by PageRank are similar, which differ from the method of rank by retweets. In addition, they categorized the trending topics based on the user participation and active period. Retweeted tweets of trending topics reached an average of 1, 000 users whether the number of followers of original tweet is large or not. Comparing the number of retweet, the types of trending topics are also the common classification (Lee et al. 2011). However, Lee et al. (2011) presented other approaches for classifying the trending topics to understanding the meaning of real-time trending topics. In order to classify the trending topics, the approach called network-based classification performed rather than the classification based on text called Bag-of-Words approach. In contrast, Zubiaga et al. (2011) proposed 15 features to characterize trending topics based on the average value of occurrences of features in the tweets corresponding to a trend and the diversity of feature values all across the tweets in a trending topic.

Although the trending topics can be classified accurately an immediate way by 15 features and with high accuracy by network-based classification, as an international social media, Twitter has a shortcoming that different countries have different interesting in trending topics of each country (Wilkinson & Thelwall 2012). They exampled that other countries are most interested in trending topics in US while least interested in Indian trending topics. But India is most interested in trending topics from other countries while US least.

Therefore, they analysed the nine month tweets from UK, USA, India, South Africa, New Zealand and Australia to find out due to international commonalities and differences, the most interested main trending topics of different countries are US holidays and US national sport events. Addressing the regional problem, the method to capture the topics in well way becomes important. Aiello et al. (2013) found out that the classic topic models as Latent Dirichlet Allocation are more proper for events during occurring in narrow scope, at the meanwhile, the methods based on n-grams co-occurrence plus time-dependent boost rather suit for the broad events. Aiello et al. (2013) analysed six different topic detection algorithms and revealed the model based on n-grams is better for more complex aggregation of keywords.

Sina weibo is a Chinese microblogging which is similar social medium as Twitter. As one of the most popular social media for people to communicate to and concern about other people, Sina weibo provides a platform which has more than 500 million users (Chen et al. 2013). They attempted to classify the weibo dataset by characteristics of users that includes lifetime of accounts, tweets per user, its followers and followings, verified accounts, geographic distribution. Due to weibo published since 2009, the different accounts with different lifetime will impact their active degree of tweeting, followers and followings. Chen et al. (2013) also mentioned that Beijing, shanghai, Guangzhou and Zhejiang are the most active area for weibo users, which means geographical distribution influences the active degree of weibo users. Beyond that the most simple and direct classification is to tank the trending list provided by Sina Weibo (Yu et al. 2012). However, Wang et al. (2014) employed Latent Dirichlet Allocation (LDA) model and Ailment Topic Aspect Model to analyse the tweets in deeper and more accurate way.

Chapter 3

Trending Topics Meaning Disambiguation

This chapter contains the evaluation study in order to identify the best approach for trending topic meaning disambiguation by comparing four different information retrieval approaches, and it was evaluated by 20 postgraduate students.

3.1 Introduction

Trending Topics are made by short phrases, keywords, or hash tags and does not include any detailed information. Majority of trending topics consists of new terms or ambiguous words. If the issue keyword is a new term or an ambiguous word, it is difficult to define the exact meaning of the topic. Let's assume that "Galaxy" is one of the trending topics in certain period. Nobody can be sure that whether the current trending topic "Galaxy" is about the phone/tablet designed by Samsung or a group of stars and planet. Without reading and analysing all the related tweets of the topic 'Galaxy', it is almost impossible to fully understand the exact meaning of the certain topics. However, it is crucial to expand the concept of trending topics to disambiguate the precise meaning of trending topics.

In order to solve the issues, several researchers have investigated summarising trending topics both manually and automatically. There is a trending topics summarisation website, called 'What The Trends', which provides the interface for users to manually type the explanation of what the trending topics is about. However, there are two different types of issues in manual trending topic summarisation approach: At first, there is the lack of labour to engage the manual inspection and secondly, majority of the explanation of trending

topic contains spam and irrelevant content (i.e. advertisement), which makes the website unreliable.

In this chapter, I inspect several researches that conducted for automatic trending topic disambiguation, compared several types of approaches, and proposed new framework for disambiguating trending topics.

3.2 Related Work

In order to disambiguate the meaning of the trending topics, there are several researches conducted for summarising and classifying trending topics. Most of researchers applied Twitter trending topics as their domain. Twitter Trending Topics provides the list of top 10 trending topics, which represents the most popular topics that are discussed by users. However, Twitter do not provide any detailed information of the trending topics but it allows to search and collect the related tweets that contains the trending topic terms. Many researchers aimed to reveal the exact meaning of trending topics using related tweets.

Sharfi et al. (2010) applied phrase reinforcement algorithm to summaries related tweets of Twitter Trending Topics. Then, the author conducted evaluation for comparing hybrid TFIDF and phrase reinforcement in use of Trending topics summarising. An experiment to compare twitter summarisation algorithms was conducted by (Inouye and Kalita, 2011). They found that simple frequency-based techniques produce the best performance in tweets summarisation. Sport events are one of the popular types in twitter trending topics so Nichols summarises the sport events. All researches in trending topics summarisation applied ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, which is extremely popular evaluation method in automatic summarisation area. Those metrics are for evaluating the quality of a summary, such as the coherence, conciseness, grammaticality, or readability. However, they are not very evaluating whether the summary contains enough contents to fully understand what the trending topic is about. Some researchers examined classifying trending topics Lee proposed classifies trending topics into general 18 categories by labeling and applying machine-learning techniques (Lee 2011). Zubiaga aimed to classify trending topics by applying several proposed features and used SVM to check the accuracy (Zubiaga, 2011). However, those researches aimed to extract the abstract of twitter trending topics but not the exact meaning represent the exact meaning of certain events.

Finding successful approaches for extracting the representative and related contents of such trending topics is very crucial in trending topic meaning disambiguation. I focused on reviewing several well-known approaches in the general information retrieval research field.

3.2.1 Word Expansion - Representative Keyword Extraction

In the information retrieval (IR) area, query expansion is a widely used technology and a typical solution that enables researchers to improve the retrieval performance. According to Efthimiadis (1996), query expansion is the process of supporting the original query by adding additional terms. It is mostly used to address the mismatched terms in the IR area. The process of query expansion generally includes four steps: resources selection, seed query construction, search results review, and query reformulation (Chum et al. 2011). Most researchers perform query expansion based on either local or global analysis (Kraft & Zien 2004). Moreover, query expansion has been performed manually, automatically, and interactively (Efthimiadis 1996). The recently introduced cluster-based query expansion is more effective than the common document-based retrieval (Lie & Croft 2004). The global analysis, local analysis, and cluster-based query expansion methods are reviewed.

Global Analysis

Global analysis is one of the most common query expansion methods because it enables improvement in that area (Aly 2008). In this approach, the collected documents are well structured are analysed, and the structure is created by the term relationship. To retrieve new documents, a user extracts the terms from the above structure. In other words, global analysis involves extracting the related keywords and results using the whole document collection. To reformulate a query, the original query is identified, and the terms that are most related to it are added (Xu, Ye & Li 2004). The initial form of global analysis was term clustering (Johns 1971). Term clustering approach was focused on expanding queries by using the terms from the same cluster. There are more techniques in global analysis, including semantic indexing, similarity thesauri, and finding phrases. One of the main advantages of the global-analysis technique is that it is relatively robust and thus tends to improve the average performance. Moreover, this approach provides a thesaurus-like structure that contains high-level domain information and different types of search support. It has a drawback, however, which is a serious problem that cannot be neglected. As global analysis requires corpus-wide statistics, it inevitably consumes a large amount of computing and time resources. Moreover, the global-analysis approach cannot effectively address the mismatching issue because it does not take the query into account (Xu, Ye & Li 2004; Xu & Croft 2000). To overcome the problems of traditional global analysis, concept-based query expansion was introduced. According to Qiu and Frei (1993), the term extraction should depend on the similarity between the collected terms and the query concept, not on that between the collected terms and an original-query

term. The evaluation by Qiu and Frei (1993) proved that concept-based query expansion is better than the traditional global-analysis method.

Local Analysis

Local analysis identifies the highly related terms in the retrieved document that are close to the query. In other words, it includes the high-ranked documents extracted by an original query (Tai 2004). It enables the user to retrieve information using an initial query, without any user involvement. The basic concept of local analysis was demonstrated by Attar and Fraenkel in 1997. In their paper, the top-ranked documents for a query were regarded as sources of information for building an automatic thesaurus (Attar & Fraenkel 1997). The most widespread local-analysis technique is relevance feedback, which reformulates the query based on the relevance judgments of the retrieved documents (Aly 2008).

Relevance Feedback

Relevance feedback was introduced around 30 years ago to improve the information retrieval performance (Salton & McGill 1983). This approach has been used for image and information retrieval (Paredes, Deselaers & Vidal 2008). Relevance feedback is considered for identifying search results based on user-centred relevance judgments. This approach can be classified into three different types: explicit feedback, implicit feedback, and blind feedback.

Explicit feedback: In a simple phrase, explicit feedback refers to the feedback received from the user. It can be categorised depending on the relevance judgments based on terms or documents. First, in document-based relevance judgment, the user should directly participate in the process of selecting some retrieved documents, whether relevant or irrelevant. After that, further terms from the documents are used when a query is reformulated in the next retrieval activity (Farah 2009). Unlike document-based relevance judgment, term-based relevance judgment is considered a supporting search activity. It first allows the user to choose highly related terms that are already auto-computed. The user should indicate the relevance of the terms by specifying the grading scale, such as “not relevant,” “somewhat relevant,” “relevant,” and “very relevant” (Liao & Veeramachaneni 2010). The selected terms are used for new-query formulation. An example of this is the search assistance feature used by Google Search. The most well-known explicit-feedback algorithm is the Rocchio algorithm (Jordan & Watters 2004). Whether term- or document-based, this approach should be used until the user is satisfied with the result.

Implicit feedback: Implicit feedback attempts to infer the users’ needs based on their previously observed actions. There are several examples of users’ observable behaviours,

such as the technique that they use, the time that they spend viewing a document or browsing a page, or noting their page-scrolling actions. After this process, feedback is given by the system (Kelly & Belkin 2001). Two examples of this are DirectHit, which ranks documents in the order of the number of people who viewed each document, and Surf Canyon browser extension, which draws the results based on the number of clicks on an icon and the amount of time spent viewing a document (Farah 2009).

Blind feedback: Blind feedback is also known as pseudo-relevance feedback or automatic local analysis. This approach allows the user to automate the manual part of explicit feedback by assuming that the top "n" documents in the results set are highly relevant. In this approach, general retrieval is implemented to search for the initial relevant-document set, and the top-"n"-ranked documents are considered the most relevant objects (Yu et al. 2003). In the end, relevance feedback is performed. Therefore, it helps improve the performance without any extension (Billerbeck & Zobel 2004). Real-time query expansion (RTQE) is a variant of the pseudo-relevance feedback. It consists of additional-query-terms suggestion when the user is typing the words (Farah 2009). RTQE is already being implemented in the search engine market, such as Google or Yahoo. For example, the suggested terms in Google are displayed in a list of recommended words (White & Marchionini 2007). For further addition, real-time query expansion carries out keyword expansion in real time rather than expanding the query from the real-time information. Fortunately, most of the blind-feedback techniques have worked well. Some researchers have proven that they work better than the global-analysis methods (Xu & Croft 2004). Moreover, the blink-feedback technique completes work automatically so that the users do not have to assess anything. One of the advantages of relevance feedback in local analysis is that it is relatively efficient in expanding the query based on the top-ranking documents. For example, if a user establishes the acceptable and correct relevance judgments, relevance feedback provides high-level performance. Moreover, it saves computer and human resources. Relevant information is utilised to renormalise the initial query by adding or discarding terms.

Local Context Analysis

Local-context analysis combines local and global analyses. It uses the passages and concepts in the global approach and applies these to a set of documents in local analysis (Benammar, Hubert & Mothe 2002). It was proposed by Xu and Croft in 1996. The concept of this approach is chosen from the top-ranked documents. Compared to the relevance feedback methodology, local-context analysis takes a much shorter time to collect the term and noun phrase collection frequencies (Xu & Croft 2004).

3.2.2 Named Entity Recognition

Named Entity recognition (NER) is one of the popular information extraction approaches, which allows extracting the predefined real-world entity or words, including as the title/name of person, location, or organisation. This approach helps all entities in the sentence or phrase so it helps people to understand the topic, which the author would like to talk about on the sentence/phrase.

NER can be worked by two main approaches, which are Rule-based approach and Statistical-based approach. Rule-based approach applies manually hand-written grammar-based linguistics, which are added by experts/linguists. Statistical-based approach applies Machine Learning (ML) techniques to extract named entities. We will now deal with the detailed previous work in the methods of named entity recognition, including rule-based named entity recognition and statistical named entity recognition approach.

Rule-based Named Entity Recognition

Rule-based Named Entity Recognition can be called as Linguistics approaches. This is because it applies the manually written linguistics itself in order to extract the named entity. Rule-based approach was traditionally proposed to obtain the higher prediction. Grishman firstly developed one of successful rule-based NER systems in 1995. The system is developed with predefined named-entity dictionary, which includes title or name of persons, cities, countries, organizations, and places. The set of rules in this system was predefined those named-entity as a text. Various rule-based NER system was developed and used for almost 20 years. However, rule-based entity recognition has fatal disadvantages. Since most of all rule-based NER system require the manually written Named entity dictionary, which is completed by the human experts (for this system, linguists). Finding highly educated and experienced linguists and providing all the related cost are very difficult. Moreover, it is almost impossible for one or two linguists to define all required grammatical knowledge of languages and convert those to computational words.

Statistical-based Named Entity Recognition

Statistical–based NER approach usually applies various Machine Learning techniques, including Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM) and Support Vector Machine (SVM). Also, it requires a large amount of annotated training data.

Hidden Markov Model: Hidden Markov Model (HMM) is one of the statistical models, which it models the pattern based on the hidden parameter. For example, it can describe as ‘Pattern Recognition’. Generally, regular Markov model allows observers to check the status

regularly so state transition probability is the only parameter that can be used. However, Hidden Markov Model is based on the outputs. Each status has possible output tokens based probability distribution. It is not possible to check the order of status based on the order of created tokens. Since it can be checked output but the status flow, the model is called as 'Hidden Markov Model (HMM). Therefore, HMM has been used in various research fields, such as Natural Language Processing, Speech Recognition, or Optical Character Recognition. For last 15 years, HMM has been dominated in Natural Language Processing and Speech Recognition. As Named Entity Recognition is a research field in both Natural Language Processing and Information Extraction, it is necessary to review the usage history. HMM has been worked based on the regular observation and labeled sequence. It is inevitable to prepare large amounts of training set. The basic idea of HMM is very simple so it is easy to implement and understand. Moreover, it uses the positive data only so it is very easily scaled.

Maximum Entropy Markov Model: Maximum Entropy Markov Model (MEMM) is also a well-known advanced conditional statistical sequence model. MEMM can be called as 'Conditional Markov Model'. The model is a graphical model, which merges the advantages of Hidden Markov Model and Maximum Entropy model. It has been known as the most convenient model to extract the named entity. As mentioned before, HMM was the most successful approach in the last 15 years but MEMM has received a lot of attention these days. Compared to the HMM, MEMM provides the increased order in choosing features to represent the observations. It is very successful in using domain knowledge to extract the required tokens. Moreover, while HMM requires applying the forward-backward algorithms in training, MEMM estimate the parameters based on the transition probabilities. Therefore, for the efficiency of cost and time, MEMM is very good approach for training all the data and tagging the features. The model has been proved that it provides increased recall and greater precision than any other NER approaches. However, several researchers pointed out that it has the bias issues in labeling.

Conditional Random Field: Conditional Random Field (CRF) is a statistical modeling approach, which is mainly used in pattern matching and natural language processing field. It is usually used in predicting structured data. While most prediction approaches uses the label of a certain sample for predicting, CRF applies content itself. In natural language processing, including named-entity recognition, CRF has been used in predicting label sequences for input data by using linear chain. The field has all necessary benefit of using MEMM but does not have to deal with labeling bias issue. Since CRF is undirected linear model, which is applying calculating the conditional probability, it is able to use as alternative approach rather than HMM. Even though it requires highly cost and time in computation the data, it offers higher-order in modeling long-range dependencies.

3.2.3 Topic Modelling

Topic modelling is the approach to discover the topics in the group of documents. The topic can be summarized as the abstract. The modeling approach enables people to find what the author/writer are talking out. This approach is very useful in reducing the processing time to summarize or discover the topic. Topic modeling approach has been used in several research fields, including search engine, machine learning and natural language processing. Topic modelling approach was initially proposed in late 1990s. Probabilistic Latent Semantic Indexing (PLSA) can be the most popular model in early topic model. After that, David Blei has been proposed Latent Dirichlet Allocation (LDA), which enables to extract the topic the most common approach for topic modelling. Since the LDA has been proposed, there are many extensions on LDA, such as Dynamic Topic Model (DTM), Correlated Topic Model (CTM), and Author Topic Model (ATM).

Latent Semantic Indexing (LSI was proposed by Hoffman in 1999. It statistical technique for the analysis of two-mode and co-occurrence data. Latent Dirichlet Allocation is a model that enables groups of observations to be explained by unobserved groups that explain why some parts of the data are similar. Most time series modeling approaches are based on continuous data, however, topic models are designed for categorical data. Dynamical Topic Model is to use state space models on the natural parameter space of the underlying topic multinomial, as well as on the natural parameters for the logistic normal distributions used for modeling the document-specific topic proportions. Correlated topic model (CTM) is a hierarchical model of document collections. CTM are based on the words of each document from a mixture model, which are shared by all documents in the collection

3.3 Data Collection

For trending topics meaning disambiguation, it is necessary to collect trending topics on Twitter and tweets related to those topics. Twitter provides an API (Application Programming Interface) that allows developers or researchers to crawl and collect the data easily. Through this API service, I collected twitter trending topics in 3 years (until 30th June, 2014).

3.3.1 Trending Topics

Twitter monitors all users' data and detects the popular trending topics that most people are currently discussing about. The detected popular trending topics are displayed on the service 'Twitter Trending Topics'. This trending topic service is located on the sidebar of Twitter interface by default so it is very easy for users to check the current trending topics

and discuss about it. It provides top 10 trending topics in real time. Hence, I have collected those top 10 trending topics per hour using Twitter API. In total, I have collected 105354 unique trending topics in 3 years. Trending topics in Twitter consist of short phrases, words, or hashtags. Twitter never provides any detailed explanation of trending topics so it is very difficult to identify the meaning of trending topics until you have a look related tweets of those topics. For example, when a missile destroys Malaysian Airlines, the trending topics were 'Malaysia Airlines', 'Malaysian', etc. It is almost impossible to realise what happened to the Malaysia Airlines by only checking the trending topics. In order to reveal the exact meaning of each trending topic, I need to collect not only the trending topics, but also the related tweets of those topics.

3.3.2 Related Tweets

The goal of this research is finding novel method to disambiguate the exact meaning and content of trending topics. To achieve this goal, it is necessary to collect the appropriate related tweets of a specific trending topic. The related tweets should not contain the contents that are irrelevant. If the trending topic is 'Malaysia Airline' which is about a missile attack happened on July 18th, I should not collect the related tweets about missing Malaysia Airline occurred on March 8th. It is extremely important to distinguish the tweets that are related to specific trending topics. Twitter API provides the tweet/search crawling service that allows users to collect the tweets by using the search query. The concept of tweet/search service is same as the search engine. Users can search the tweets that contain the search keyword. The search results contain detailed information of each tweet, including content, username, location, created date-time, and etc. We used this created date-time to extract the appropriate tweets for the trending topics. As I collect the top 10 trending topics in an hourly basis, I search and collect the related tweets that users upload in last one hour. For example, when 'Malaysia Airline' is on the trending topics list at 8pm, I search and collect the related tweets that users upload in last one hour, 7pm to 8pm. This collecting approach prevents irrelevant tweets

3.4 Methodology

As mentioned before, the aim of this chapter is to find novel method for disambiguating the exact meaning of the trending topics in Twitter. We focused on examining whether the methods are sufficient to extract the appropriate contents that represent the specific trending topics. We experimented with four different methods that are applied in topic-sense

disambiguation research field: Key Factor Extraction, Named Entity Recognition, Topic Modeling, and Automatic Summarisation. The philosophies behind these four methods are very different, but each has been shown to be very effective in the information retrieval area.

3.4.1 Key Factor Extraction

The first selected method for twitter trending topic sense disambiguation is the key factor extraction by applying numerical statistic. There are several key factor extraction approaches that are aimed to find the most important keywords in the document by calculating the importance weights of each word.

TF (Term Frequency) weighting is a classic key factor extraction technique for automatic determination of term relevance. The term frequency in the given tweets gives measure of importance of the term within the particular document. TF weighting is a classic approach but still widely used in Information retrieval area. TF can be determined the exact values in various ways, such as raw frequency, boolean frequency, logarithmically scaled frequency, and augmented frequency. We used raw frequency calculation, which is the most classical approach. The TF weighting tf(t,d) can be calculated by counting the number of times each term occurs in a document. However, like most English sentences do, tweets include several common words, such as ‘the’ or ‘a’. Assume I calculate TF weights for all terms in documents including those extremely common terms. Since the term ‘the’ is too common, the result will point the term ‘the’ as the most important word. In order to solve this issue, I eliminate all stop-words from tweets. The list of stop-words I used is based on the ‘Full-Text Stopwords in MySQL’. After removing those stop-words, I applied TF weighting to identify the important terms in the related tweets of each specific trending topic.

3.4.2 Named Entity Recognition

Named entity recognition (NER) is widely used for labelling the name of objects in documents. It labels sequences of terms, which are about the name of objects, such as person, organisation, or location. By recognising named entities, it can be easy for people to identify what kind of subject/topic the document is discussing about. We applied one of the most popular Named Entity Recognition approach, Conditional Random Field (CRF) sequence model. CRF-based NER are investigated by Stanford NLP lab and it is widely used as a standard NER technique. CRF is a type of probabilistic sequence model, and it is applied for sequential data labelling. The basic idea of CRF sequence model is as follows. Assume X is a random variable over data sequences to be labelled, and Y is a random variable over corresponding label sequence. The nodes in the model are separated into two different sets,

X and Y. A conditional distribution p(Y|X) with an associated graphical structure will be modeled.

CRF-based NER models are trained by the official sources such as dictionary or WordNet. The applied CRF model for this study is trained on the CoNLL English training data. For extracting the named entities in related tweets, I applied this trained 4 classes CRF model that contains the entity information of person, location, oragnisation and misc.

3.4.3 Topic Modelling

Topic Modelling is the approach that discovers the abstract topics in the multiple documents. The discovered topics consist of a cluster of words that frequently occur. LDA (Latent Dirichlet Allocation) is the most successful approaches in topic modelling area. The concept of LDA can be explained with the following example. If multiple documents are randomly mixed over various types of topics, the topic can be characterized by a distribution over words. LDA is very different from the traditional Dirichlet-multinomial clustering model. Like many other clustering models, traditional clustering model does not allow a document to being clustered with a single topic. However, LDA has three levels, and notably the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics. We used LDA approach in Mallet Topic Modelling tool for training and testing the representative content extraction, with all parameter set to their default values.

3.4.4 Automatic Summerisation

Automatic Summarisation was introduced for people to save the document reading time by providing a summary that retains the most important points of the documents. There are two main approaches, extraction and abstraction, in automatics summarisation. According to the evaluation conducted by Inouye and Kalita, most extraction approaches produced better performance; especially SumBasic had the highest scores in ROUGE metrics. SumBasic is a frequency based summarisation system, which uses the following algorithm. First, it calculates the probability distribution over the words in the input data. For each sentence in the input, assign a weight equal to the average probability of the words in the sentence. Then, select the highest scored sentence that contains the best probability word. For each word in the chosen sentence, update the probability. If the desired summary length has not been reached, go back to the first step. In this research, I applied SumBasics to extract the summary of related tweets of each specific trending topic.

3.5 Evaluation Set-up

After collecting data and selecting the approaches to apply, I conducted evaluation to find the most successful approach in twitter topic sense disambiguation. Unlike other studies, I do not focus on the readability or conciseness of extracted content, but examine which approach can extract the most relevant and representative content of a specific trending topic. As mentioned before, I collected 105,354 twitter trending topics and tweets related to them in 3 years. With this in mind, I randomly selected 100 different trending topics and related tweets for each topic. Then, I selected four different types of information retrieval approaches, including Key Factor Extraction (TF), Named Entity Recognition (CRF sequence model), Topic Modelling (LDA), and Automatic Summarisation (SumBasics). Each selected approach disambiguates the sense of trending topics in their own way by using the related tweets.

Table 3.1 shows the example contents extracted from the related tweets of a trending topic 'Susan Powell'. Those contents display the result of applying four different information retrieval approaches. The trending topic 'Susan Powell' was on the list in 7th February 2012. It was about the following news. Josh Powell, husband of missing Utah woman, killed himself and his two young sons in Washington house fire. He was a murder suspect of his wife. You can find specific information about the topic from the result of KFE with TF.

Table 3.1 The contexts extracted of a specific topic by applying four algorithms

Approaches	Extracted Contents
KFE with TF	Susan, Powell, Josh, powell, Utah, sons, woman, killed, Cox, boys, doubts, fate, missing, death, PollyDad, Charlie
NER with CRF	[Susan/P, Candlelight/P, Washington/P, Cox/P, Cheyenne/P, Utah/L, Miller/P, Itâ/P, Powellâ/P, Charlie/P, Husbandâ/P, City/L, Powell/P, Brandon/P, WEST/L, Wash/P, VALLEY/L, Josh/P, Mommy/P, Marc/P, Denise/P, Candlelight/L, Klaas/P, Kids/P, Dad/P, CITY/L, West/L, Valley/L, Tacoma/P]
TM with LDA	susan, family, lovely, watched, flips, middle, black, husband, children, afternoon
AS with SumBasic	josh powell Any doubts about Susan Powell's fate should be dispelled in lieu of Josh Powell's homicidal binge.

* The trending topics for this example is 'Susan Powell'

For evaluating the performance, extracted contents of each 100 trending topic is assessed by 20 postgraduate students in Computing and Information Systems. All students are trained

by attending 2 hours workshop for this evaluation. In the workshop, students are encouraged to understand that the evaluation is focusing on the quality of the extracted contents for trending topics, not the readability of the contents. For this evaluation, I developed the evaluation system as can be seen in the figure 3.1. Figure 3.1 displays the user interface after a student logged into the system.

In the workshop, participants are asked to use the evaluation system in the following order:

1. Choose one of the 100 topics on the top left section, 'Trending Topics'

2. After selecting a specific topic, the related tweets will be shown on the top right section, 'Related tweets'

3. Click any related tweet to read. The content of each tweet will be displayed on the middle section. By reading those related tweets, get the point what the trending topic is about.

4. After fully understanding the specific topic, grade based on the content extracted by four different approaches in the sense disambiguation area.

In the evaluation, the grades are given based on Likert scale (from highest to lowest) 1,2,3,4, and 5. (Check the figure 3.2). The meanings of those five grades are as follows: 1=strong agree, 2=agree, 3=neutral, 4=disagree, 5=strong disagree. By using this grade, student give grades for each extracted contents. If a student agrees with only the output of the KFE with TF, but strong disagree with 3 other results, they can give grade '2' to KFE and '5' to the content of all 3 other approaches. The evaluation was successfully conducted in 10 days.

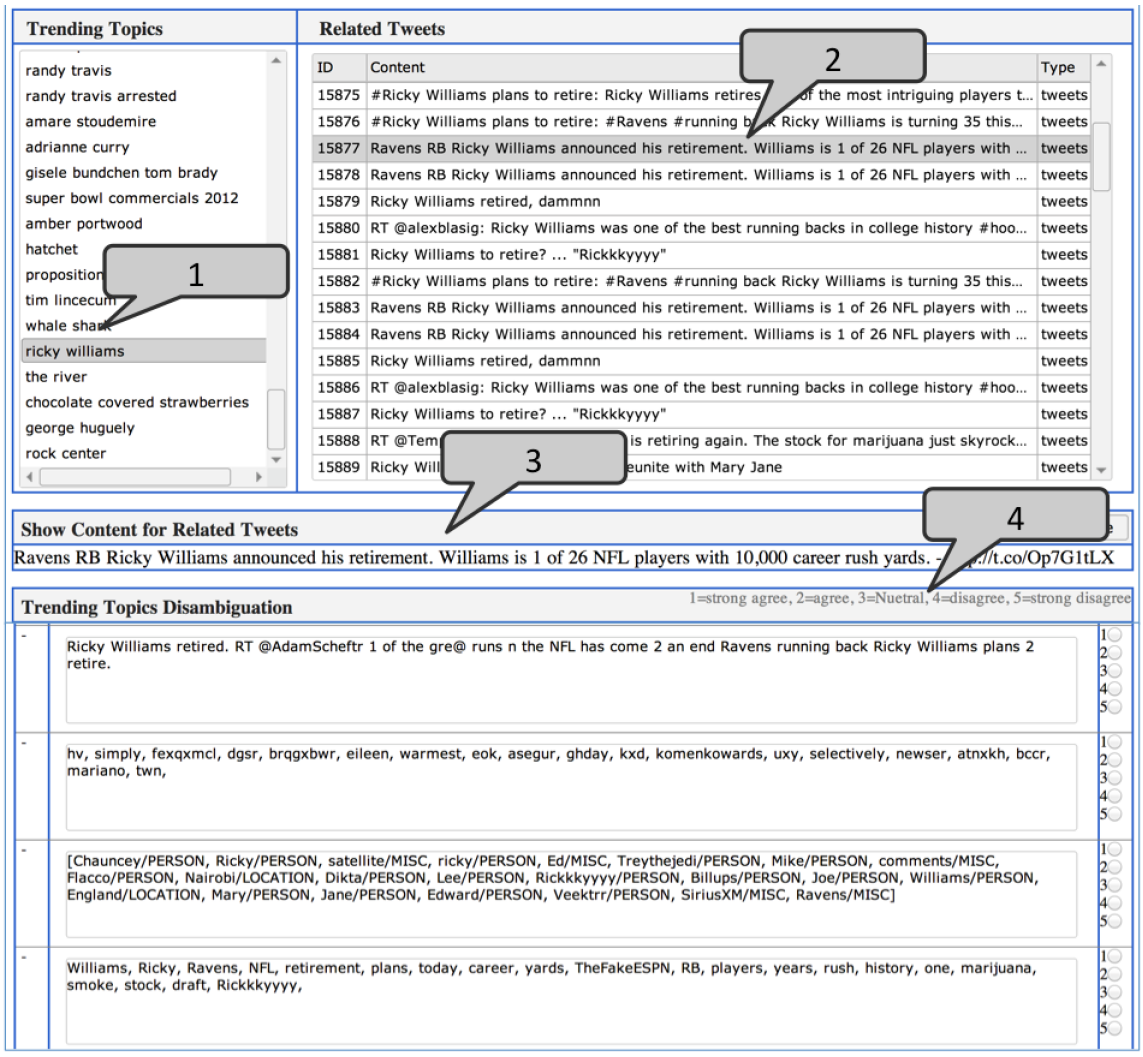

Fig. 3.1 The human evaluation system interface

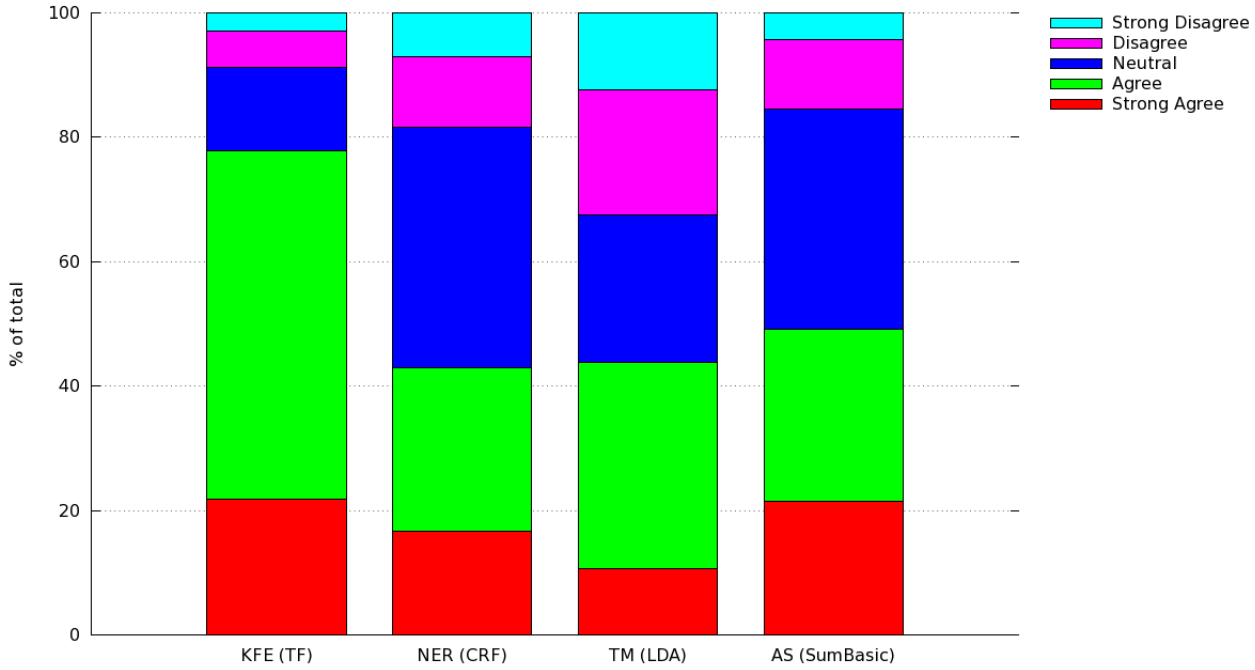

Fig. 3.2 The grade distribution for four different approaches

3.5.1 Evlauation Results

First, I begin the analysis of result with the following question: How much score each approach receives in average? As can be seen in the table 3.2, the average score for all 4 selected approaches are between 2(agree) and 3(neutral) level.

Table 3.2 Average Likert Score for each approaches

Approaches	KFE with TF	NER with CRF	TM with LDA	AS with SumBasic
Average score	2.12	2.66	2.90	2.49

It seems all 4 approaches are generally acceptable for twitter trending topics sense disambiguation, and the key factor extraction receives the highest grade among those approaches. However, the average score is not enough to define the successful approach. A graph in figure 3.3 shows that the distribution of the responses on each approach. The graph clearly indicates that only few participants (less than 10%) strong disagree with the output of sense disambiguation for all evaluated approaches, except topic modelling. The participants roughly understand the meaning of twitter trending topics with the extracted contents of all chosen approaches.

As I have seen in the average scores, Keyword Factor Extraction (KFE) got the highest (almost 80%) positive responses (for both strongly agree or agree) and it can be clearly seen from the distribution as well. It is a quite interesting result that the participants provided positive responses with the output of KFE that is extracted based on the classical term frequency weighitng technique. However, I found that the contents from Named Entity Recognition (CRF) and Automatic Summarisation (SumBasic) are not clear enough, since the neutral responses took the biggest percentage in those approaches.

Based on the result shown in the above figure, it seems KFE is the good approach to extract the representative contents of the twitter trending topic. However, the KFE result is not fully covered, as there are few amount of neutral and negative response. About this issue, I analysed and found that there is a high correlationship between all four approaches.

As you can see from the figure 3.4, 3.5, 3.6, and 3.7, KFE has inverse correlationship with all 3 different approaches in its negative responses (disagree-4 and strong disagree-5). This indicates that other approaches can be used as a substitute of KFE, when the extracted contents are not satisfied. However, this would require an analysis of twitter trending topics to find proper conditions for interchangeability and this is left for our future work.

Likert Scale	1=strong agree, 2=agree, 3=neutral, 4=disagree, 5=strong disagree

Fig. 3.3 The grade distribution for four different approaches

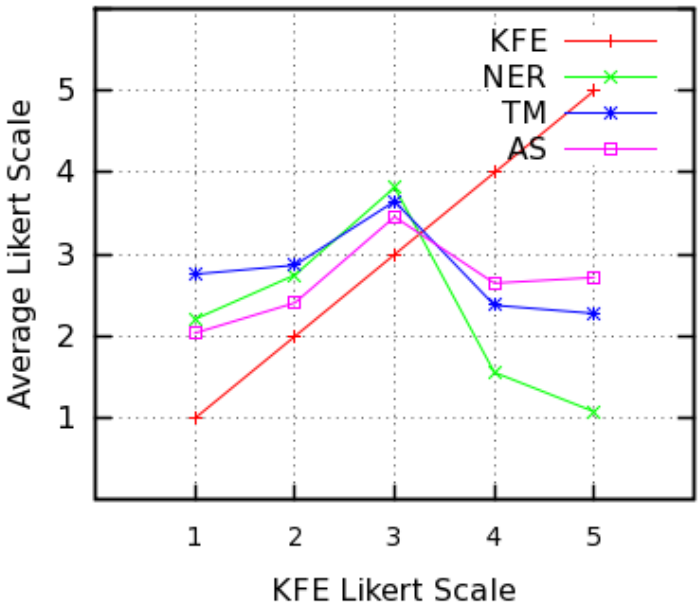

Fig. 3.4 The grade correlation analysis among four different approaches KFE based

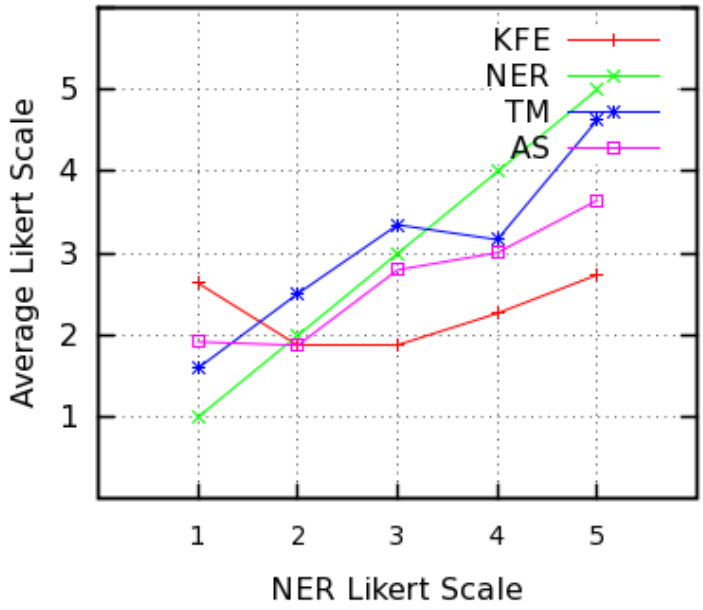

Fig. 3.5 The grade distribution analysis among four different approaches NER based

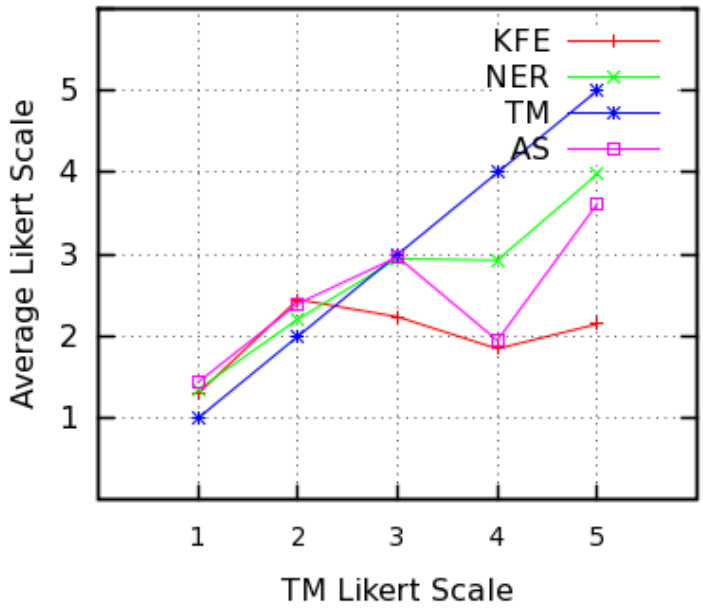

Fig. 3.6 The grade correlation analysis among four different approaches TM based

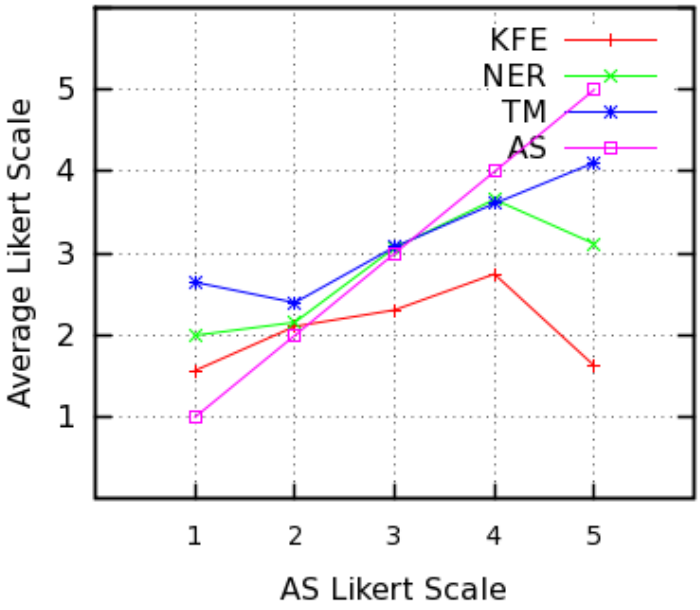

Fig. 3.7 The grade correlation analysis among four different approaches AS based

3.6 Implementation

The following list represents the database for this research designed for storing the raw trending topics, data processed by four different information retrieval approaches for human evaluation. The database for this project is designed as follows:

- Table: tb_twt_keyword
  - id: primary key, the identification number generated by auto increment function.
  - keyword: twitter trending keyword
  - rank: rank of the twitter trending keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - country: country of the twitter trending keyword
  - local_time: local time at collection
  - date: collect time
- Table: tb_twt_relatedTweets
  - id: primary key, the identification number generated by auto increment function.

  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - tweet_id: the identification number of the related tweet given from twitter api
  - tweet_content: contents of the related tweet
  - tweet_date: uploaded time of the related tweet
  - retweet_count: the number of re-tweet of the related tweet given from twitter api
  - favorite_count: the number of favorite of the related tweet given from twitter api
  - date: collect time
  - tb_twt_relatedTweet_user_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweet_user table

- Table: tb_twt_relatedTweet_user
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_relatedTweet_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweets table
  - user_id: the identification number of the twitter user given from twitter api
  - user_name: the identification username of the twitter user
  - user_screenName: the screenname of the twitter user
  - user_location: the location of the twitter user uploaded tweet
  - user_followers_count: the number of followers of the twitter user given from twitter api
  - user_friends_count: the number of friends of the twitter user given from twitter api
  - date: collect time

- Table: tb_twt_relatedNews
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - news_content: contents of the related news
  - news_date: uploaded time of the related news

  - source: the source of the collected related news
  - date: collect time

- Table: tb_keyword_meaning_disambiguation
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - keyword: twitter trending keyword
  - content_tweet_kfe: key factor extraction result of the related tweets
  - content_tweet_ner: named entity reconiser result of the related tweets
  - content_tweet_tm: topic modeling result of the related tweets
  - content_tweet_as: automatic summarisation result of the related tweets
  - content_news_kfe: key factor extraction result of the related news
  - content_news_ner: named entity reconiser result of the related news
  - content_news_tm: topic modeling result of the related news
  - content_news_as: automatic summarisation result of the related news
  - content_combined_kfe: key factor extraction result of the related news and tweets
  - content_combined_ner: named entity reconiser result of the related news and tweets
  - content_combined_tm: topic modeling result of the related news and tweets
  - content_combined_as: automatic summarisation result of the related news and tweets
  - date: collect time

- Table: tb_meaning_disambiguation_user_evaluation
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - tb_evaluation_users_id: foreign key, the identification number that enables to connect with tb_evaluation_users table
  - approach: meaning disambiguation approach type
  - user_likert_score: likert score of the approach given by the user

  - date: collect time

- Table: tb_evaluation_users
  - id: primary key, the identification number generated by auto increment function.
  - username: username of the user
  - password: password of the user
  - date: collect time

The summary of database design can be found in the Appendix A.

3.7 Conclusion

As mentioned before, the goal of this chapter is finding the most successful method to retrieve the representative contents for twitter trending topics sense disambiguation. In order to achieve the goal, I first collected the trending topics and tweets related to them. Then, I applied four different information retrieval approaches, including key factor extraction, named entity recognition, topic modelling, and automatic summarisation. We conducted human experiments with 20 postgraduate students. Based on results reported in the chapter, the statistical key factor extraction approach, a classical term weighting technique, provides the highest performance in retrieving the most representative contents for trending topics sense disambiguation. However, topic modelling does not work well on finding the topic words from those real-time events information. As mentioned before, I present the result of the first human evaluation in online trending topic sense disambiguation. We hope the research is the step forwards improving the performance for any researches using trending topics as a data.

Chapter 4

Trending Topics Relevance Identification

Chapter 5 proposes new framework that identifies the relevance of trending topic to a certain target domain, including an individual or organisation. By idenfiying the personalised relevance, people or organisations are easy to get how much the certain trending topic is related to them.

4.1 Introduction

There is no doubt that social networking services (SNS), one of the real time online communication services, have been received much attention recently (Sakaki, Okazaki & Matsuo 2010). This service is used by millions of people around the world to share information or their interests with their family members, relatives, friends, etc. SNSs ask for the users' status, such as through questions like "What are you interested in?" or "What's happening with you?" As soon as the users give an answer to this question or update their status, all the people connected to them in the SNS can see their status. Therefore, the main reason for the extraordinary popularity of SNSs is that they enable the users to communicate with others in an efficient way, and reflect the users' real-life behaviours and interests (Tirado et al. 2011). SNSs change not only the way that people communicate with one another but also the speed of sharing information. There are two reasons for the increase in the speed of sharing information. Unlike other online communication services, SNSs provide push-based information. For example, while the e-mail is like a letter placed by a person in another's mailbox that the latter can open when he/she wants to, SNS can be likened to the user tapping another person's shoulder and forcefully placing a message on the latter's hand. Second, SNS messages are broadcasted to all the people linked to the senders while e-mail or instant messages are sent only to the address/addresses specified by the sender. In other words, the recipients are less limited in SNSs than in the other online communication services (Barasa

2010). Due to these two characteristics of SNSs, people can share information much faster than ever.

As the speed of communication flow has been increased by SNS, a large amount of information exists on the Web, and because humans are social beings and are thus intensely interested in what others are doing, there are those who want to see what information people are looking for. Many search engines and Websites did not pass up this opportunity. For example, Google, Yahoo, and Twitter are actually providing a new trending service regarding the topics that people are searching most frequently for and are most interested in. According to some researchers, among the existing trending services, Google provides the most powerful service, called "Google Trends." Researchers also claim that Google Trends has a list base of highly searched terms and phrases. Therefore, it is an efficient service that identifies the search trends in real time (Rech 2007; Kwak et al. 2010). However, there is no service that visualises the relationship between a trending topic and a person or organisation. While most researchers neglect this area, it is very crucial for both individuals and organisations to respond to trending topics. This is because certain trending topics may have a significant impact on a person or organisation. If people or organisations know the relevance between a certain hot issue and themselves, the latter will be able to identify the opportunities and threats that such trending topic may present to them. It will thus help them prepare for such opportunities and threats as soon as possible. On the other hand, if people/organisations do not know the relevance of a trend on them, the trending topics will merely become a source of fun and useless gossip. This well represents the motivation of this study.

The aim of this research is providing adapted/personalised relevance identification service by identifying relevance between trends and a target object, such as an individual or organisation. The research presents the results of the investigation of the visualised relevance of the trending topics on the target organisation/individual.

4.2 Related Work

To develop the system that is personalised with a certain target object, such as individual users or organisation, they always need to provide the digitalised domain. Fortunately, most activities for both individuals and organisations are saved in assortment of digital information recently (Kushmerick & Lau 2005). There are several kinds of digital information management systems for both individuals and organisations. Most information management system is well-structured and categorised. Moreover, those systems are centralised storage, by covering almost all activities of a target object (Voida, Harmon & Al-Ani 2011).

4.2.1 Personalised Domain

Email

E-mail was initially designed as an electronic asynchronous communication tool (Whittaker & Sidner 1996). Since recently, most people store their activities and information in their respective e-mail accounts. Many researchers look into how people actually manage their information via e-mail. According to Boardman and Sasse (2004), e-mail is the one of the most successful individual information management tools. The set of messages in an e-mail account reflects the life of the account owner because many of his/her life's highlights are stored therein, either in the e-mail inbox or in the archives section (Voida, Harmon & Al-Ani 2011).

Structure: There is no doubt that e-mail provides rich information to individuals. To classify all these information, e-mail account owners manage their e-mail using folder hierarchies. Folders are constantly reorganised, some folders are created, and other folders are emptied and deleted (Brutlag & Meek 2000). Recently, a system was created and implemented in which once an e-mail account owner creates a hierarchical structure of folders, most e-mail services assist the user by providing the function of automatic message classification. There are many examples of the e-mail classification system, but the main tool for it is text-based classification (Kiritchenko & Matwin 2001). Most of the text-based approaches classify messages into pre-defined categories based on their texts. POPFile is one of the most successful tools for automatic e-mail classification (Kamens 2005).

Blog

Blog was introduced as an information-sharing technology that updates and exchanges ideas/interests (Kavanaugh et al. 2006). It is generally maintained by an individual who shares his/her thoughts through such technology, a process called "self-disclosure." (Ko & Pu 2011). This technology allows people to observe an individual activity by accessing each post (Higgins, Reeves & Byrd 2004). According to Fitzpatrick (2007), blogs are used as both knowledge-sharing and personal-work/information storage spaces. Most blogs have a specific topic, such as food, travel, entertainment, or movies, which coincide with the administrators' areas of interest (Nakajima et al. 2005).

Structure: The users update their ideas and interests through a process called "blog post." Blog posts are usually arranged in chronological order. Moreover, most blog posts are categorised by predefined directories. Some users classify blog posts in the pre-defined categories by manually clicking on or selecting them. Many blog services automatically classify posts using various features, such as titles, tags, and descriptions. A typical example

of this technology is tag-based blog classification, which has received much attention from researchers of late (Sun, Suryanto & Liu 2007).

Knowledge Management

In organisations, document management is becoming increasingly important because the system structure in the industrial field is becoming increasingly complicated. Most document classification activities are done by people in the organisation. To classify the documents in the organisation, people use the tacit knowledge rather than explicit knowledge. With the tacit knowledge, it might cause problems because people are not usually aware of the importance or exact meaning of the knowledge. Many researchers have been worked on transforming tacit knowledge into explicit knowledge. There are several widely-used methods to transform knowledge, such as cognitive mapping (Rodhain 1999), Ikujiro Nonaka's model (Erden, Krogh & Nonaka 2007). Almost all organisations engage in knowledge management, which involves the values from the organisation's assets (Juang, Lin & Kao 2008). Many organisations have adopted a computer-based system that supports the integration of decision and knowledge management. This system is called "knowledge management system" (KMS) (Cheng, Lu & Sheu 2008). A KMS is a virtual repository of relevant information for organisational knowledge workers (Dalkir 2005). KMS was originally designed as a proce ss of applying a systematic approach that includes knowledge acquisition, storage, and dissemination. After that, KMS came to be used by organisational workers and provided feedback from the users (Richardson, Courtney & Haynes 2006).

As mentioned earlier, the KMS process involves four activities. First, explicit and tacit knowledge is acquired. Explicit knowledge is put in either paper or electronic format while tacit knowledge is organised from people's minds. Second, the collected knowledge is stored in electronic-document form in a virtual repository and is accessed and shared by the employer. In the end, the knowledge is utilised by the workers, and the feedback is updated by the users. The process flow of KMS are well declared in Roknuzzaman, Kanai & Umemoto 2009 and King 2009. Benefits: KMS presents several benefits to organisations. Here, five benefits will be reviewed. First, KMS encourages users to manage new ideas and information in the free flow of data. Second, it tends to manage their customers well by providing accurate information and instant response. Third, it expands the organisation's revenue as both products and services are marketed much faster than ever. Fourth, it encourages the rate of employee retention by letting the organisation to recognise the value of knowledge. Further, organisations are led to create inducement programs by motivating their employees to share information and knowledge. Finally, it streamlines the organisation's operations, and

the organisation's costs are reduced as the system eliminates the redundant or unnecessary processes (Batten 2008).

4.2.2 String Comparison and Relevance

The prior works on string comparison and matching methods are briefly reviewed. String comparison has long been a research topic in computer science. It is regarded as a method of string matching that enables the system to make decisions using the actual content flow (Tan, Brothrton & Sherwood 2006) In other words, string comparison has been defined as the process of measuring the similarity between strings. It is performed in many pattern-matching and Web search areas (Cormode & Muthukrishnan 2005). As string comparison has a long history, various methods of it have been introduced.

Edit-distance Method

String edit distance is one of the most popular notions in string comparison (Ristad & Yianilos 1998). The similarity between two strings can be specified based on the distance between them. Assuming that A and B are strings, the more similar they are, the lesser the distance between them. The edit distance between two strings is regarded as the minimum number of editing activities needed to transform one into the other, including insertions, deletions, substitutions, and in some cases, character transpositions (Cormode & Muthukrishnan 2002). As edit distance has been widely used in information theory and computer science, there are various kinds of distance measures in edit operations.

Damerau-Levenshtein Distance: The initial form of edit distance method was introduced by Damerau (1964) and Levenshtein (1966) (Yarkoni, Balota & Yap 2008). The Levenshtein distance was the first to be introduced. It could perform only three basic editing operations: the insertion and deletion of a single character and the substitution of one for another. Damerau extends the Levenshtein distance by allowing one more edit operation: transposition. Therefore, the Damerau-Levenshtein distance is the standard simple edit distance metric (Brad 2007). The distance between "ELEPHANT" and "RELEVANT" is 3. First, "r" must be inserted at the beginning of the word (Elephant -> Relephant). Then "p" must be substituted with "v" (Relephant -> Relevhant). Finally, "h" must be deleted (Relevhant -> Relevant). There is no way to lessen the edit operations between the two words. This metric also covers the uppercase and lowercase characters. For instance, the distance between the two strings "Apple" and "apple" is 1 because only the substitution of the uppercase "A" with the lowercase "a" is required. he initial motivation of the Damerau-Levenshtein measure was

calculating the distance between two strings to correct the spelling errors therein. It has been used in biology, however, for validation between DNAs.

Needleman-Wunsch distance: The Needleman-Wunsch distance is an extended version of simple edit distance, such as Damerau-Levenshtein. The general idea of the Neeleman-Wunsch distance is providing the global alignment of two sequences steadily and recursively. This algorithm consists of two phases: the scoring and Trace-back phases (Xia & Dou 2007). First of all, it produces two matrices of real numbers, called "scores." The scores of ranged characters are regarded as the similarity matrix (Markov & Kalinin 2010). The details of this approach are as follows. In such approach, the maximum score of the best path is computed, and the score is placed in the score array (Lunter et al. 2008). As there are several mismatches and a large gap between the two sequences, the penalties were set at (1), (-0.5), and (-0.5), respectively. The bold line indicates the best alignment path that is 4.

Jaro-Winkler Distance

The Jaro-Winkler distance is the approach that measuring similarity between two strings. It is originally used by the U.S Census Bureau for comprising people's name. The initial metric was introduced by Matt Jaro and it was redeveloped by Bill Winkler. This approach is mainly used in the field of duplicate detection and record connection (Cohen & Sarawagi 2004). The Jaro-Winkler distance between two strings is based on how similar two strings are. If the Jaro-Winkler distance for two strings is higher, those two strings are very similar. According to the Cohen, Rvikunar & Fienberg (2003), this approach is well-designed and intended for short strings such as personal first or last name. The distance metric is normalised between 0 and 1; 0 represent that there is no similarity and 1 signify the exact match.

TF/IDF Distance

The TFIDF (term frequency inverse document frequency) distance is the weighting method that is usually performed in the field of information retrieval. The approach is focused on evaluating the importance of a word to a document in a collected domain. The importance is based on the number of a word appearance in the given document, however, that is balanced by the word frequency of the word in the collected domain. There are a lot of variants of TFIDF that is widely-used in search engines, such as Google or Yahoo Search. Most search engines are used this approach in scoring and ranking the relevance of document on a given initial query. TFIDF can be also used in stop words filtering in the several kinds of research files, such as text summarisation or classification. Summing the TFIDF weight of each query term is regards as the simplest function in ranking field. Almost all advanced ranking

functions are variants of this simplest function (Salton & McGrill 1983). This algorithm is composed of two phases as follows: TF (Term Frequency) and IDF (Inverse Document Frequency). First of all, it is necessary to count the number of a term occurrence in a given document. This number of term appearance is often normalised to reduce the issue of the document's length. If the length of 'A' document is longer than 'B', 'A' document has more chance to obtain much higher number of term appearance rather than 'B' document. In this case, the actual importance of each document will be shown incorrectly. Therefore, Term frequency method measures the importance of the term t in a given document d. The TF method defined as tf(t,d), which is tf = (the number of term appearance in a given document/ the total number of words in a given document). For example, support there is a document that contains 200 words. If the term 'paper' appears 20 times in that given document, the TF weight is (20/200) = 0.1.

However, there is a significant issue if there is common term as a query term, such as 'the'. If the term 'the' appears 60 times in the given document, the TF weight is (60/200) = 0.3. The TF weight of 'the' is much higher than term 'paper' but it is not a good idea to distinguish term 'the' is more relevant to that document than 'paper'. When the common term is shown, there are lots of chances to get the wrong results. Therefore, in this case, IDF (Inverse Document Frequency) is necessary to perform support TF. The IDF is a popular measure of a word's importance. The method of IDF weight calculation is by dividing the total number of documents in the collected domain by the number of documents that contains the query term.

Suppose there are 20000000 documents in the collected domain. If a term 'the' appears 20000000 times and a term 'paper' occurs 2000 times. Each IDF weight would be measured as log (20000000/20000000) = 1 for a term 'the' and log (20000000/2000) = 4 for a term 'paper'. As TFIDF score is calculated by combing the TF weight and IDF weight. The TFIDF weight of each terms are specified as follows: term 'the' – $0.3 \times 1 = 0.3$, term 'paper' – $0.1 \times 4 = 0.4$. Therefore, the TFIDF is based on a high TF and a low IDF weight. (Jones 1972) Many researchers indicate that the TFIDF term weight is providing good performance in the relevance decision making field (Wu et al. 2008; Hammond 2010).

4.3 Methodology

In this research, the methodology that was employed in this research can be divided into four phases, as follows: (1) trending topic collection; (2) relatedkeywords extraction; (3) target domain management; and (4) relevance calculation. First, trending topics were collected from Google Trends, a service that shows the top 10 searched keywords. It is hard to define

the exact meaning of a trending topic using only one Google Trends keyword. The related keywords extraction phase involves a method of showing the exact meaning of a topic by extracting related keywords from microblog and news. To provide the personalised impact of a trending topic on a target object, not only the trending topics are required but also the personalised target domain. The target domain should cover the entire breadth of knowledge of a target object. It should be well structured and digitalised. Such phase will be followed by the impact calculation/visualisation phase.

4.3.1 Trending Topics Collection

The first phase involves the collection of the trending topics that show what people are currently most interested in. Without any real-time dataset, it is hard to identify what the most popular topic is at the moment. Fortunately, most search engines and Websites addressed this issue. For example, Google, Yahoo, and Twitter provide the new service of showing the list of trending topics in Google Trends, Yahoo Buzz, and Twitter Trending, respectively. In this research, Google Trends was chosen as the trending topic collector. Google Trends (2011) displays the list of the top 10 fastest-rising search terms based on hourly data from Google Search. The search-terms indicate that topics people are interested in and looking for. Yahoo Buzz also provides an hourly list of the top 10 hot search terms. However, it is evident that Google Search is currently the most popular Web search engine (Kwak et al. 2010), for which reason, Rech (2007) indicated that Google Trends most effectively provides the highly searched terms and phrases. Thus, Google Trends was chosen as the trending topic collector in this study so that more accurate results would be obtained. Unlike Google Trends, Twitter provides the top 10 trending topics based on tweets or messages in Twitter. The flow of Twitter, however, is influenced by "big mouths" like celebrities. The service provided by Google Trends, however, is based on the fastest-rising search terms in Google Search. Search results are affected not just by big mouths but also by the general users. Therefore, Google Trends was chosen because of its objectivity. First, the page that displays the list of fastest-rising search terms was collected from Google Trends. As Google Trends updates the list hourly, the page was collected once per hour. After collecting the page that contains the list of top 10 search terms, 10 keywords were extracted from the page.

4.3.2 Related Keywords Extraction

Even though the top 10 trending topics per hour were collected, there was a significant issue with regard to deriving the exact meaning of each trending topic. Ambiguity occurs when the exact meaning of a trend topic is obtained using each trend from Google Trends. For

instance, assuming that “Apple” is one of the fastest-rising search terms in Google Trends, most people may think that the keyword “Apple” is an American multinational corporation that sells computer materials. The keyword “Apple,” however, may be related to the fruit or farm thereof. It is obvious that drawing the exact meaning of each trending topic from Google Trends will result in ambiguity. Therefore, it is necessary to expand a trending topic by extracting several related keywords. There is then a need to decide where and how to extract the related keywords. Before doing so, the characteristics of the trending topics should be identified. All trending topics are based on time-related information because they represent the most popular topics at the moment. As mentioned earlier, Google Trends displays the list of fastest-rising search terms as real-time trending topics. Therefore, the related keywords must be extracted from a service that provides real-time publishing, such as microblog and Internet news services. According to the Sakaki, Okazaki & Matsuo (2010), the most important characteristic of microblog is real-time nature. In addition, people read Internet news to view the real-time updated contents. Using microblog and Internet news, the appropriate related- keywords that can help obtain the exact meaning of a Google Trends keyword can be extracted. If related keywords are extracted from general documents published at any time, without any specific period, semantically related keywords will be extracted, not keywords that are related to the trending topic. In this research, to extract the appropriate related keywords from microblog and Internet news, documents related to a Google Trends keyword were first searched. Twitter and Google News were chosen as the microblog and Internet news service, respectively. As it is necessary to extract only documents related to a particular trending topic, the latest three pages about a Google Trends topic were collected. The collected documents from Twitter and Google news are pre-processed by Stemmer and Part-Of-Speech (POS) tagger that helps user to catch the nouns. This is because most related keywords are comprised of nouns. After extracting nouns from the collected documents, I use Term Frequency (TF) method to find the most relevance nouns on a Google Trends keyword. By using TF weight, it first counts the number of times each certain term appears in the collected document. In this phase, I assign to each term weight based on the number of occurrence of the term in the document. However, the system could be found much higher number of term occurrences in the longer documents rather than shorter one. Therefore, TF weight will be defined by dividing the occurrence count of a certain term by the total number of words in the given document. Then, each term weights are sorted in ascending order. The top p keywords would be the highest related keywords. The best number of related keywords will be chosen in evaluation part. As mention earlier, the system collected related documents by using Google Trends keyword. represents the related documents from microblog, Twitter. represents the related documents from the internet news.

4.3.3 Target Domain

As finished the trending topic collection, the second goal of this research is how to manage the target domain. In order to calculate the relevance between trending topics and a target, it is essential to collect not only trending topics, but also a target's information and activity. The target would be an individual or organisation. Each trending topic might be related to a certain part of a target domain. Then, how can I identify the relevance between a certain trending topic and each part of target domain? The best way to identify the target domain is to obtain the digitalised document management system that contains all up-to-date information and activities of a target object. Moreover, the document management system should be well-structured and categorised.

The typical examples of digitalised and well-structured document management system are email, blog, and KMS; both email and blog is the well-managed individuals' activities and information, however, most organisations are using KMS as document management system, which categorised each section by using folder. The way to categorise the document is usually decided by people so that it might not be subjective. However, the relevance will be viewed by people who classified that way so that it might be not a problem. To collect up-to-date information regularly, it is necessary to use document mornitoring system that detects and stores new or updated information.

What can be the most appropriate style of target domain? Since each individual's email or blog is concentrated on a few narrow topics, it might not be shown the relevance in various trends. KMS provides hierarchical structure with appropirate documents but it might contain private information and not available to public. In this research, I applied web monitoring system (Web-Mon KMS) in order to constract the well-structured KMS, which contains up-to-date activities and information as a target domain (Part, Kim, and Kang 2003). Australian government websites are selected for the target domain in this research. This is because it contains several departments that can be covered broad area, and categorised into well defined hierarchical structure. All documents will be monitored and collected from each department websites regularly as can be seen in figure 4.1.

The website monitoring system are working as following steps:

- identify the website url for constructing target domain
- monitor the web documents from the identified websites regularly with checking the freshness
- store webpages and folders in database
- preprocess (stemming and POS tagging) the stored web documents

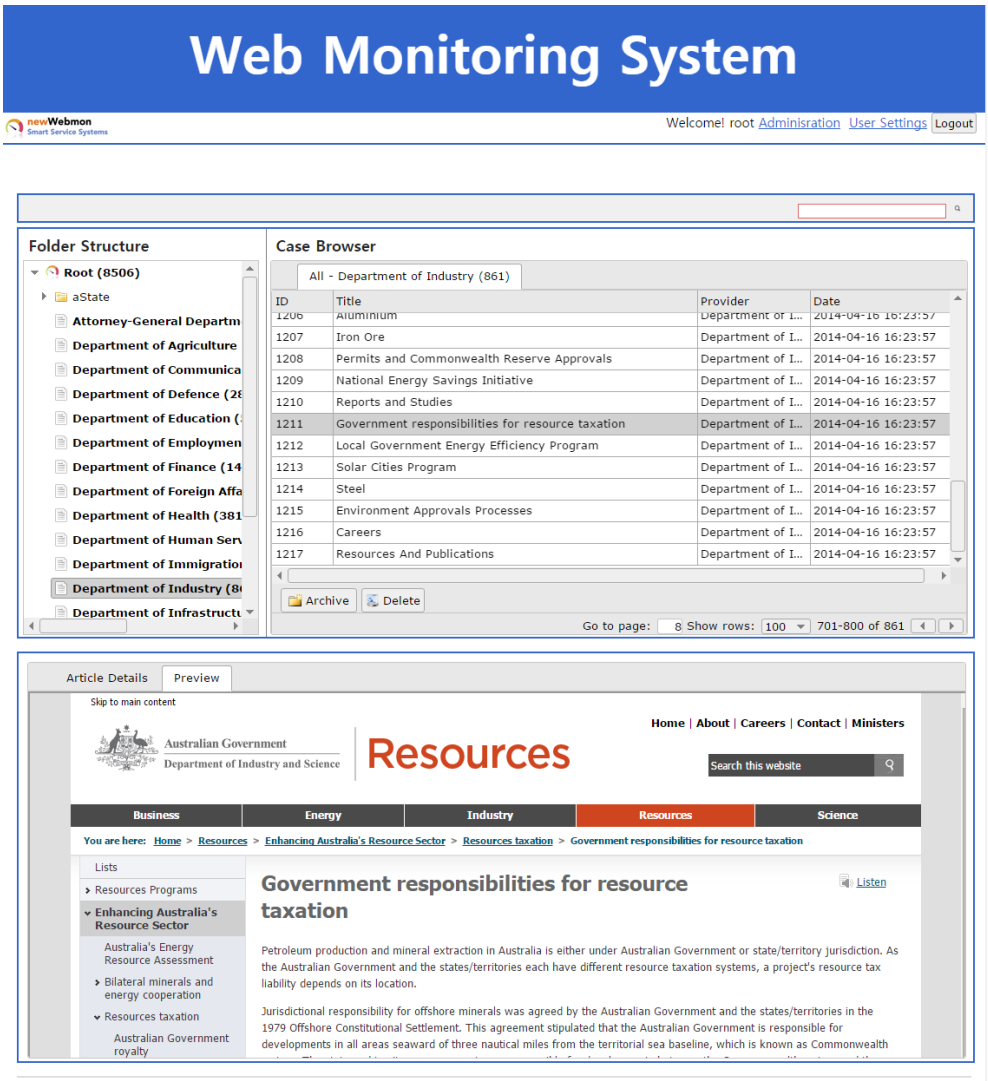

Fig. 4.1 Target Domain -Australia Government

The target domain for this research have different folders based on the structure of departments websites that are declared by Australia Government (http://www.australia.gov.au/).

For example, assume the ‘deparment of health’ website are monitored and all those web documents from the website are stored in ‘department of health’ folder. The web monitoring system collects all web documents that include text-based data form without images, videos, or any multimedia resources. All collected web documents are pre-processed by stemming and POS tagging to extract only nouns, like I did in trending topics documents. The reason why I extract only nouns is to make easier to calculate the relevance between trends and a target domain. More details will be explained in next phase.

4.3.4 Relevance Identification

The final goal of this research is calculating relevance between the collected trending topics and the target domain. In this research, trending topics are collected from Google Trends, Twitter, and Google news. The target domain are comprised the combination of various department websites in Australia government. The way to collect both trending topics and target domain has been mentioned before. In order to identify the relevance between trends and the target domain, I applied the Term Frequency Inverse Document Frequency approach that is the most widely-used approach in calculating similarity and string comparison field. It is usually used by search engines that need to rank a document's relevance given a user query. This aspect is very important in this research as well.

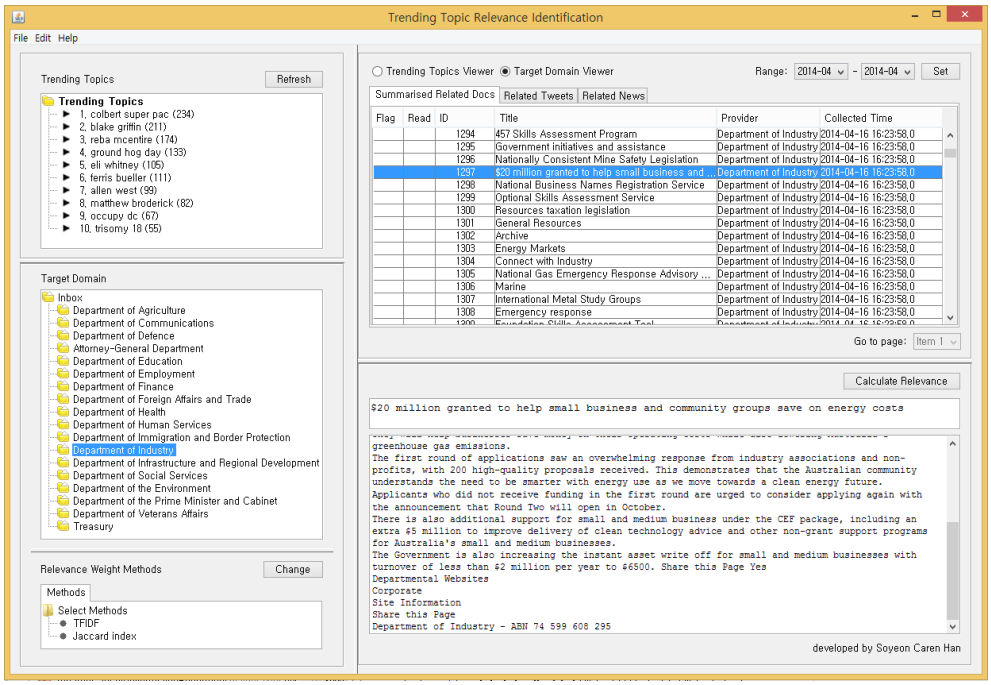

Fig. 4.2 trending topic relevance identification

To explain how the TFIDF method is applied to this research clearly, the example will be provided as below. As mentioned earlier, the set of trending keywords includes one Google Trends keyword and four related keywords from Twitter and Google news. Suppose a Google Trends keyword is 'sushi' and the related keywords are 'roll-ups', 'udon', 'enchanted' and 'food', which extracted from the related information (tweets and news). What the system wants to obtain is the relevance of each documents to the set of trending keywords. First, the system removes the documents that do not contain all five trending keywords, including 'sushi', 'roll-ups', 'udon', 'enchanted', and 'food'. After that, the system counts each number

of terms. If a document contains 200 words and the number of term ‘sushi’ occurrence in that document is 5. Following the previously defined formulas, the ‘Term Frequency (TF)’ for ‘sushi’ is (5 / 200) = 0.025. However, there is an issue with term ‘food’ because it is very common word in those food blogs. For example, if the number of term ‘food’ occurrence in the same document is 50, the weight is (20 / 200) = 0.1. If I use only Term Frequency (TF), it is incorrectly emphasizes ‘food’ rather than ‘sushi’. Therefore, it is necessary to use Inverse document frequency (IDF) as well. Assume the total number of documents in the target domain is 1000000 and ‘sushi’ appears in 10 documents. From the defined formulas, the weight of ‘Inverse Document Frequency (IDF)’ is measured as log(1000000/10) = 5. Therefore, the Term Frequency Inverse Document Frequency (TFIDF) weight of term ‘sushi’ is $0.025 \times 5 = 0.125$. On the contrary, if the ‘food’ appears in 1000000, the TFIDF weight is $0.1 \times 1 = 0.1$. Therefore, ‘sushi’ is much important than ‘food’ in this case. The weight in TFIDF is calculated by both a high TF (in one document) and low DF of the term in the whole target domain. Hence, TFIDF tends to filter out very common terms. In this research, the relevance weight of a document is computed by summing TFIDF weight for each 5 keyword, including one Google Trends keyword and four related keywords. Moreover, the relevance weight of a folder is calculated by summing TFIDF weight for each document. Finally, the total relevance weight of a target domain is identified by combining all documents’ TFIDF weight.

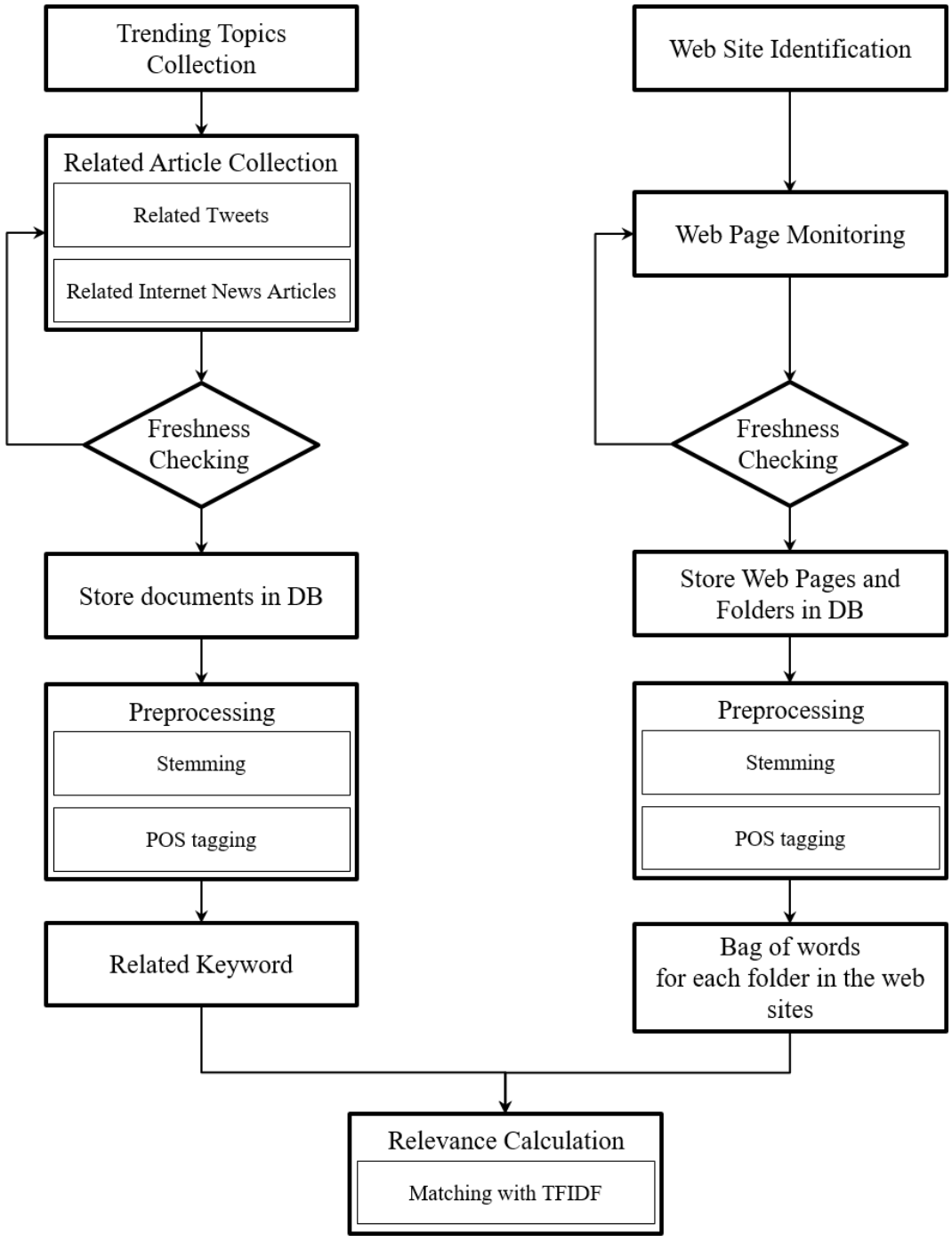

Fig. 4.3 Flow diagram for trending topic relevance identification

4.4 Evaluation Set-up

Evaluations of the proposed system were carried out in order to examine the success of the method. With this in mind, I collected data for evaluating the proposed method. First, to extract trending trending topics, I crawled Google Trends keywords for a period of 195 days, approximately over 5 months. As described in the introduction, I obtained 17559 unique topics. Secondly, in order to reduce the ambiguity of the trending topics, I extracted several related keywords from Twitter and Google News hourly. The target domain is the combination of different countries' food blogs, which were collected from Google search. In the target domain, there are 4 continent categories (e.g. Asia), 14 area categories (e.g. East Asia) and 26 country categories. We crawled 22933 documents. We collected only the blogs, which are shown in the first page of Google Search. Each data set contains one Google Trends keyword, several related keywords, date, and relevance weight. We calculated not only each target's relevance weight, but also the relevance weight of each document and each category.

4.5 Evaluation Results

In the first part of evaluation, we explain the reason why we extracted several related keywords. To do this experiment, we extracted 10 related keywords for each Google Trends Keyword, and calculated their relevance weights.

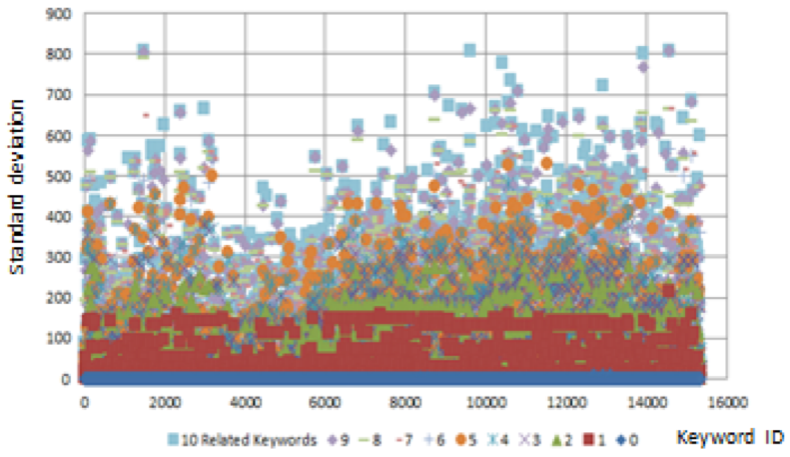

Fig. 4.4 Relevance weight based on the number of related keywords

Figure 4.4 displays the relevance weights for the number of related keywords. First, when we did not extract any related keyword, most relevance weights are almost 0, which can be

seen in the blue line at the bottom. If the relevance weights are almost 0, it may be very hard to distinguish which social issue is highly related to a target. On the other hand, you can clearly see the difference when we extracted at least one related keyword. This result proves the importance of the related keywords extraction.

Figure 4.4 well represents the importance of the related keyword extraction but it is not easy to see how the relevance weights are changed. In this chapter, we provide Figure 4.5 which displays the standard deviation value of relevance weights for the number of related keywords. In Figure 4.5, the x-axis represents the number of related keywords. It shows the more related keywords are extracted, the easier it will be to distinguish between documents.

Figure 4.6, which shows the standard deviation, median and average of relevance weights for the number of related keywords. As can be seen in the graph, the more I obtain the related keywords, the higher will be the standard deviation, median and average weights. According to this result, it will be easier to distinguish among documents if I extract more related keywords are extracted.

Next, I consider the appropriate number of related keywords. In Figure 4.5, you can see the gap between each standard deviation is dwindling. This result might show the proper number of related keywords to identify the personalized relevance of trending topics to a target object. There are two reasons why I would like to obtain the most appropriate number of related keywords. First, I have to consider the time needed. We collected related articles from Twitter and Internet news hourly; Tweets are almost 90 and news articles are almost 10. It depends on the number of articles that people uploads in an hour.

Extracting over 10 related keywords may not require a long time, but it does require a great amount of time to calculate the personalized relevance of a Google keyword and over 10 related keywords to a target. Secondly, the number of the related articles is limited. If I extract over 10 related keywords, some keywords might not be really related to that trending topic. In other words, some keywords may just be very general words that have no relationship with a Google Trends trending topic keyword. Therefore, for these two reasons, it is necessary to obtain the suitable number of the related keywords.

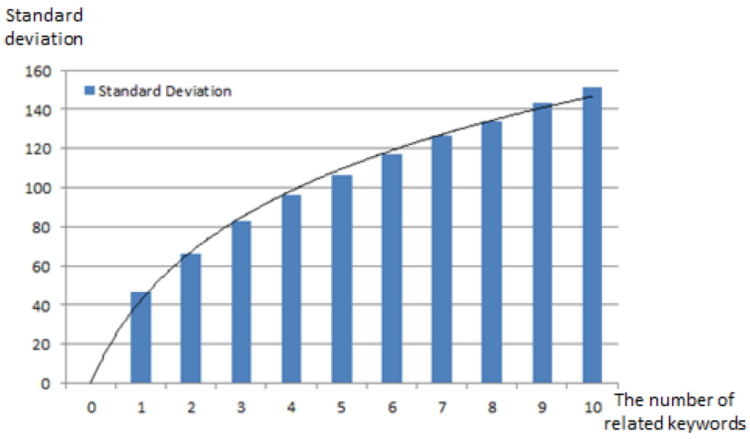

Fig. 4.5 Standard deviation for TFIDF relevance based on the number of related keywords

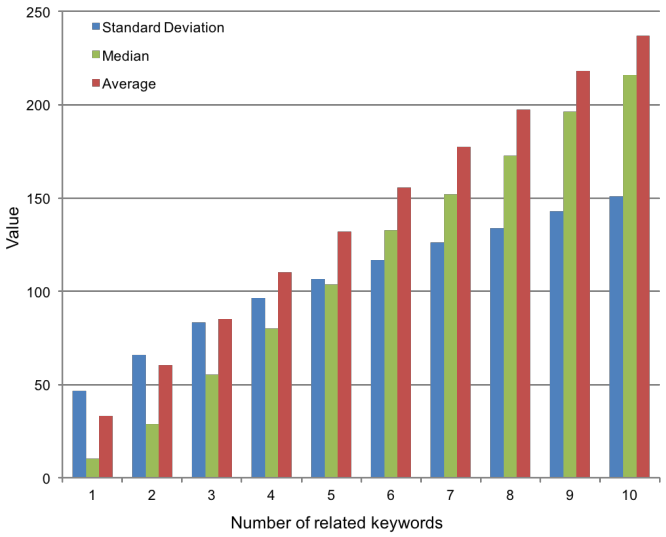

Fig. 4.6 Standard deviation, Median and Average for TFIDF

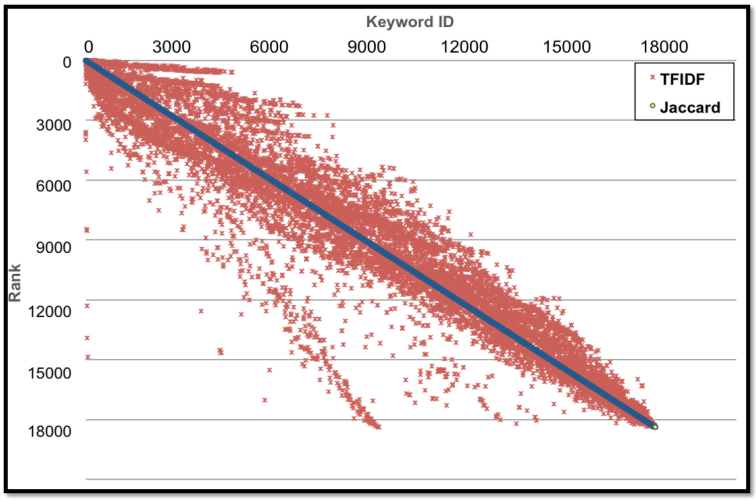

Fig. 4.7 Relevance Weight Comparison using TFIDF and Jaccard

In this research, I used TFIDF as a primary approach to calculate the relevance weight of social issues to a target. TFIDF is a good approach for calculating relevance weights and it is usually used for ranking relevance weights in most search engines. However, it has never been used in this area before. Therefore, we conducted an experiment to prove the efficiency of TFIDF by comparing it with another relevance weight approach, Jaccard.

As can be seen in figure 4.7, in order to find a similarity of trend between TFIDF and Jaccard weight, we ranked each keyword on the basis of its applied relevance value. Then we ranked in ascending order these social issue keywords that are applied TFIDF method and compared them with the rank of same keywords that are applied in Jaccard. In general, similar trends are observed in both of two methods, TFIDF and Jaccard. Therefore, we can obtain the similar relevance weight regardless of relevance weight approach. For future work, it might be good to propose new relevance weighting approach that will suitable to this project.

4.6 Implementation

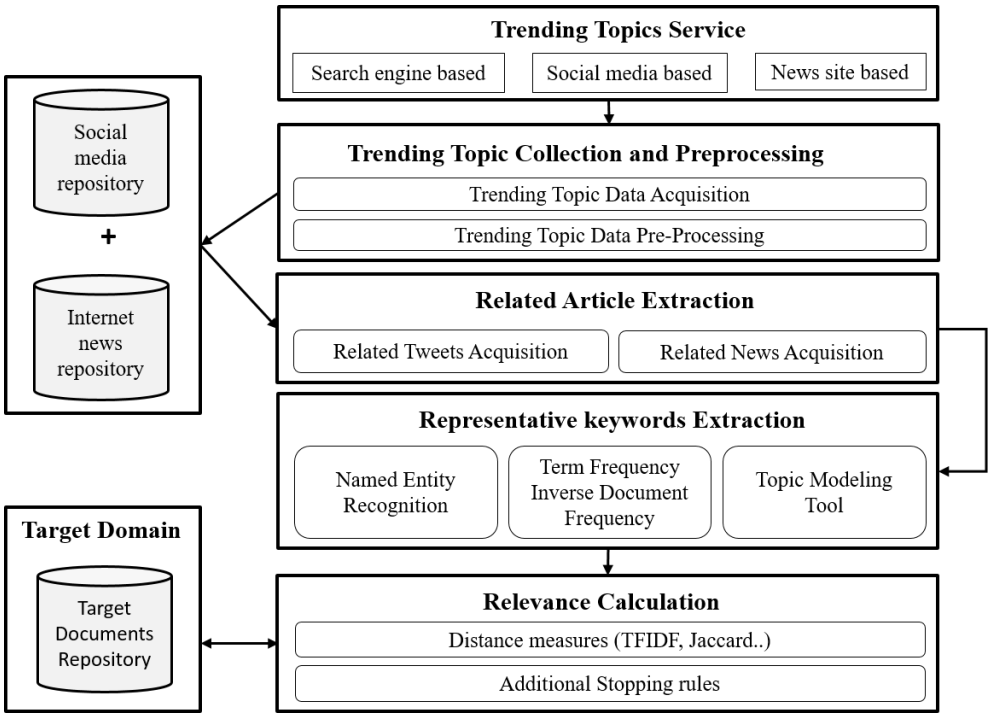

Fig. 4.8 System Architecture for trending topic relevance identification

4.6.1 System Operation

The system has been implemented to test the methods that mentioned. The proposed system operates the following order: First, the system collects the trending topics hot keywords from Google Trends. The collected keywords store in the database. After that, in order to solve ambiguity of each keyword, the system search and extract related documents from microblog and internet news. The collected related documents stored in the database. Since I need related keywords from these documents, the system extracts only nouns as most keywords are comprised of nouns. To extract only nouns, the system performs stemming, POS tagging, and the term frequency method. After collecting the set of trending topics, the system should identify the relevance between a trending topics and a target domain. The documents from a target domain are already pre-processed by stemmer and POS tagger. Finally, the system calculates the relevance between the group of trending topics and a target domain by using the Term frequency inverse document frequency method. The calculated relevance stored in the database. An overall flow of the proposed system can be seen in figure 4.8.

The database for this research is designed for storing all detailed information of trending topics from Google trending topics analytics services with the related information. Meaning

of each stored trending topic is disambiguated by using related information and store in the database. The rest of tables are for storing target domin monitored by Web-Mon and relevance identification weight of monitored/structured target domain. The database for this project is designed as follows:

- Table: tb_ggt_keyword
  - id: primary key, the identification number generated by auto increment function.
  - keyword: google trending keyword
  - rank: rank of the google trending keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - country: country of the google trending keyword
  - local_time: local time at collection
  - date: collect time
- Table: tb_ggt_relatedTweets
  - id: primary key, the identification number generated by auto increment function.
  - tb_ggt_keyword_id: foreign key, the identification number that enables to connect with tb_ggt_keyword table
  - tweet_id: the identification number of the related tweet given from google api
  - tweet_content: contents of the related tweet
  - tweet_date: uploaded time of the related tweet
  - retweet_count: the number of re-tweet of the related tweet given from google api
  - favorite_count: the number of favorite of the related tweet given from google api
  - date: collect time
  - tb_ggt_relatedTweet_user_id: foreign key, the identification number that enables to connect with tb_ggt_relatedTweet_user table
- Table: tb_ggt_relatedTweet_user
  - id: primary key, the identification number generated by auto increment function.
  - tb_ggt_relatedTweet_id: foreign key, the identification number that enables to connect with tb_ggt_relatedTweets table

  - user_id: the identification number of the google user given from google api
  - user_name: the identification username of the google user
  - user_screenName: the screenname of the google user
  - user_location: the location of the google user uploaded tweet
  - user_followers_count: the number of followers of the google user given from google api
  - user_friends_count: the number of friends of the google user given from google api
  - date: collect time

- Table: tb_ggt_relatedNews
  - id: primary key, the identification number generated by auto increment function.
  - tb_ggt_keyword_id: foreign key, the identification number that enables to connect with tb_ggt_keyword table
  - news_content: contents of the related news
  - news_date: uploaded time of the related news
  - source: the source of the collected related news
  - date: collect time

- Table: tb_keyword_meaning_disambiguation
  - tb_ggt_keyword_id: foreign key, the identification number that enables to connect with tb_ggt_keyword table
  - keyword: google trending keyword
  - content_tweet_kfe: keywords extraction result of the related tweets
  - content_tweet_ner: named entity reconiser result of the related tweets
  - content_tweet_tm: topic modeling result of the related tweets
  - content_tweet_as: automatic summarisation result of the related tweets
  - content_news_kfe: keywords extraction result of the related news
  - content_news_ner: named entity reconiser result of the related news
  - content_news_tm: topic modeling result of the related news
  - content_news_as: automatic summarisation result of the related news

  - content_combined_kfe: keywords extraction result of the related news and tweets
  - content_combined_ner: named entity reconiser result of the related news and tweets
  - content_combined_tm: topic modeling result of the related news and tweets
  - content_combined_as: automatic summarisation result of the related news and tweets
  - date: collect time

- Table: tb_target_page
  - target_page_id: primary key, the identification number generated by auto increment function.
  - target_page_name: page name of the target site
  - target_page_url: URL of the target site
  - monitoring_scheduler_id: foreign key, the identification number that enables to connect with tb_monitoring_scheduler table
  - target_registerd_time: registered time of the target site

- Table: tb_monitoring_scheduler
  - monitoring_scheduler_id: primary key, the identification number generated by auto increment function.
  - monitoring_interval: monitoring time interval of the scheduler

- Table: tb_fetched_page
  - fetched_page_id: primary key, the identification number generated by auto increment function.
  - target_page_id: foreign key, the identification number that enables to connect with tb_target_page table
  - fetched_page_url: url of the fetched page
  - fetched_page_title: title of the fetched page
  - fetched_time: collection time of the fetched page

- Table: tb_page_content

  - fetched_page_id: foreign key, the identification number that enables to connect with tb_fetched_page table
  - extracted_content: extracted content of the fetched page
  - crawled_time: collection time of the crawled page content

- Table: tb_domain_foldertree
  - folder_id: primary key, the identification number generated by auto increment function.
  - parent_folder_id: parent folder id for constructing folder structure
  - folder_name: folder name
  - creation_date: created time of the folder

- Table: tb_domain_article_relevance
  - fetched_page_id: foreign key, the identification number that enables to connect with tb_fetched_page table
  - tb_keyword_id: foreign key, the identification number that enables to connect with tb_ggt_keyword table
  - relevance_weight_type: relevance weight calculation type
  - relevance_weight_value: relevance weight calculation value
  - insertion_date: inserted time of the iscalculated

4.7 Conclusion

As described in this research, the aim of this research was providing adapted/personalised relevance identification service by identifying relevance between trends and a target object, such as an individual or organisation. The outcome of conducted initial tests proved that I achieved the three primary goals: (1) how to collect the trending topic, (2) how to manage a target domain, and (3) how to calculate the relevance between the trending topic and a target domain. Firstly, it is proved in the evaluation that the system can extract the accurate related keyword from Twitter and Internet news. The advantage of extracting four related keywords is shown. However, more examination is required to conduct to know the best number of related keywords. As I constructed the virtual target domain that is well-structured and categorised, the system can identify the relevance value of each document and folder.

Therefore, the user can see which document or folder have high-relevance on a trending topic. To identify the relevance, the proposed system used TFIDF method that is widely-used in relevance identification field. The proposed system provides the relevance value for all documents and folders. In final conclusion, despite remaining some uncertainties about how the method should be implemented, it can be seen that the proposed system, personalised relevance identification system, is a valuable service. Unfortunately, it is still required to determine the full potential of the method, and how to identify the accurate relevance. Moreover, the system is currently collecting the data so that it can be conducted further examination to determine the full potential of this new system.

Chapter 5

Trending Topics Lifecycle Prediction

Chapter 6 proposes new approach for trending topic lifecycle prediction, and it proves that statistical analysis are possible to predict the future trends of trending topic lifecycle.

5.1 Introduction

By using different types of web-based services, such as search engines, social media, and Internet news aggregation sites, internet users can share and search information through the world. These services have caused a huge information-sharing paradigm shift by increasing personal information sharing and acquisition. This phenomenon, often called "the social data revolution", has resulted in the accumulation of unprecedented amounts of social data. This large amount of user created social data is like an untapped vein of gold in 21st century. Many information providers analyze their social data and provide a Trending Topics service, which displays the most popular terms that are discussed and searched within their community. Various companies, including Google, Yahoo, Baidu, and Twitter, have been providing this trending topics service for more than 5 years.

One of these services, Twitter, monitors their social data, detects the terms (including phrases and hash-tags) currently most often mentioned by their users, and publishes these on their site. Their list displays the top 10 trending topics of the moment, and displays these as part of the Twitter interface so all users can easily identify the current trending topics. Abdur Chowdhury, a chief scientist at Twitter Research Team, defined the Twitter Trending Topics as below: Twitter Trending Topics helps people understand what was happening around the world showing us that people everywhere can be united in concern around important events Trending topics are estimated to reflect the real-world issues from the people's point of view. Kwak et al (2010) demonstrated that over 85% of trending topics in Twitter are related to breaking news headlines, and the related tweets of each trending topic provides more detailed

information of news and users' opinions. Hence, being able to know which topics people are currently most interested in on Twitter, and their point of view, may lead to opportunities for analyzing the market share in almost every industry or research field, including marketing, politics, and economics?

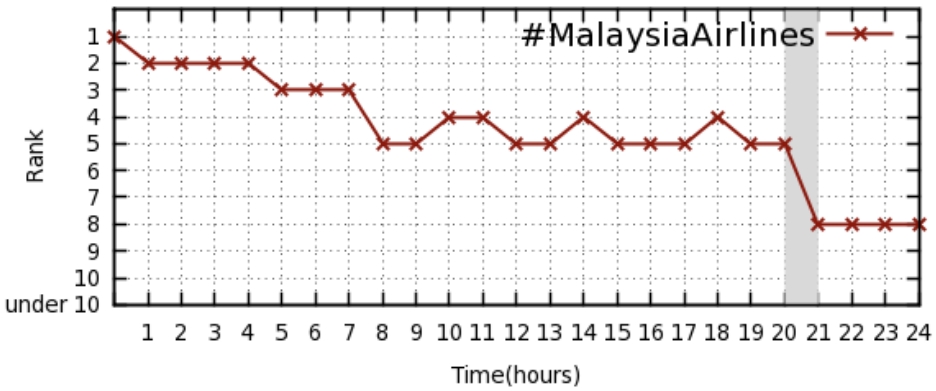

Fig. 5.1 The ranking pattern of trending topic '#MalaysiaAirlines'

The 'Trending Topics' list shows the top 10 trending topics in descending order of popularity. The lower the rank the higher the popularity, the higher the rank the lower the popularity. Based on the rank of a trending topic, it is possible to recognize the degree of current popularity of that topic. For example, on July 17th 2014, when a missile downed the Malaysia Airlines plane over Ukraine, all the 295 passengers and crews were killed in the blast and it was breaking news around the world. During this time, the topic '#MalaysiaAirlines' appeared on the trending topics list. Figure 5.1 shows the hourly rank changes of the trending topic '#MalaysiaAirlines' in 24 hours, which start from the point the topic first appeared on the trending topic list.

As you can see in figure 5.1, the trending topic has different hourly ranking changes based on people's interest change. The hourly ranking changes can be classified into three categories: up, down, and unchanged. In other words, this hourly ranking change represents the degree of change of popularity in that topic every hour, whether the people's interest in each trending topics is going up, down or staying unchanged. Predicting the trending topics hourly ranking change can be helpful to identify the influence of the topic in the near future.

However, the 'Trending Topics' list displays only limited information, including the trending topic term, its rank, and updated date and time. We used only this available information to predict the future rank changes of trending topics. Therefore, the research aim of our study is to answer the following question: "how can I predict the change of trending topics' popularity (up, down, and unchanged) by only using historical rank data?" In order to solve the problem, I proposed a temporal modeling framework using historical rank data and

machine learning techniques. At time t, the problem, predicting the future rank change FRC of a trending topic T_x, can be expressed as follows:

where f is a machine learning technique and the historical rank of n period is $[r_{t-n}, ..., r_{t-1}, r_t]$. The predicted ranking change FRC of trending topics can be classified into three classes: up, down, or unchanged. For example, let's assume that I predict the rank change of '#MalaysiaAirlines' from 20hours to 21hours (down), I can use its historical rank data from 0 to 20 hours.

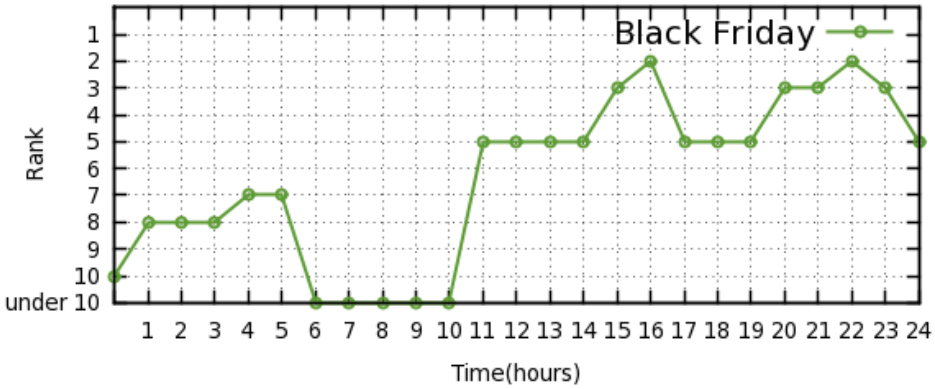

Fig. 5.2 The ranking pattern of trending topic 'Black Friday'

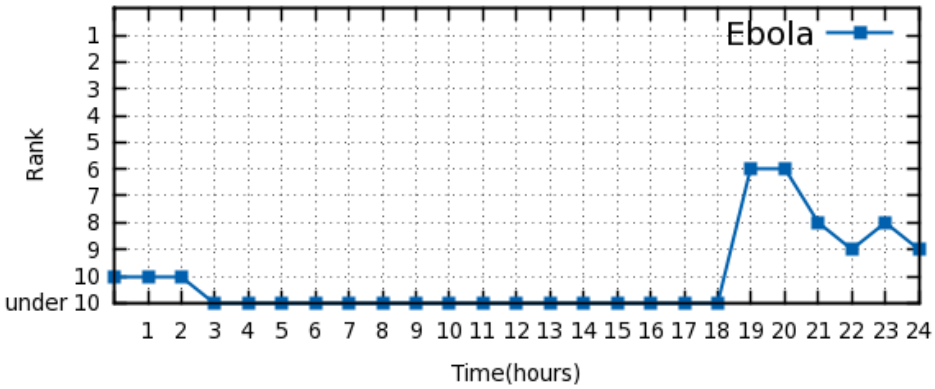

Fig. 5.3 The ranking pattern of trending topic 'Ebola'

In order to use the historical rank data for rank change prediction, there is an issue to investigate. Several trending topics tend to disappear and reappear from the trending topics list so it is impossible to know the exact rank when it disappears. Figure 5.2 and figure 5.3 well represents the example of this topic disappearance issue with two different trending topics, 'Ebola' and 'Black Friday'. It shows the 24 hour rank pattern, which represents its rank from the point the topics was first detected on the 'Trending Topics' list. The topic

'Black Friday' is a seasonal trending topic. Twitter users are talking about what they will do on Black Friday. The topic appeared around the lowest ranks before the day, and disappeared for 5 hours. It reappeared in the morning of that day. Another topic 'Ebola' relating to the deadly virus infection which raised concerns due to a resurgent epidemic in West Africa. In the initial three hours, the topic was about the first infection in Guinea and then it disappeared for 16 hours. Following that, the topic reappeared and was about another infection in Liberia. Hence, these are the same topic term but each has a different context.

Historical rank data show that almost 70% of trending topics tend to disappear and reappear later. Therefore it is important to reflect this 'disappearance and reappearance' phenomenon in the prediction model, which is related to handling missing value and window size.

First, it is necessary to handle the missing ranking value while the topic disappears. We applied four different missing value-handling approaches in methodology, and identified the most successful approaches in trending topics rank prediction in the evaluation. Secondly, as historical rank data is time-series data, it is necessary to select the optimal window size. Rather than choosing a random window size, I proposed a method to select the appropriate window size for predicting rank change of trending topics. As can be seen in Figure 5.2 and figure 5.3, the context can change while the trending topic has disappeared. We need to find the minimum length of topic disappearance hours in the same topic with different contexts, and apply it to the window size.

The contribution of this chapter are summarised as follows:

- The chapter shows that the popularity change is predictable by using only historical rank pattern.
- This work is the initial work that proposes a temporal model for predicting the future trends of trending topics' ranking behavior, and has not been reported in the literature before.

5.2 Related Work

Social media, such as Twitter, is considered a very successful tool in identifying what people are interested in and their point of view. By using the historical data in social media, called 'wisdom of the crowds', the following research has been conducted for predicting real-world events in various fields, including politics and economics.

First, there are several studies aimed at predicting political events, such as election result, using social media. One of the most representative studies was applying tweets

and news analysis for predicting election (Chung and Mustafaraj, 2011). They collected all related tweets and news that contain the candidate's name and conducted sentiment analysis. O'Connor et al. focused on determining what U.S. citizens think about each party by conducting sentimental analysis of tweets in 2010. The proposed approach was based on the simple text analysis but it considers the tweets as time-series data. UK general election was also forecasted by Franch. The author used social data from various types of social media, including facebook, twitter, youtube, and those media are classified into its media level. The prediction performance was evaluated by using ARIMA(auto regressive integrated moving average), and it achieves 0.48 and 0.83 percentage points off the real vote share. In the last four years, researches have been conducted in various countries elections using social media data, including U.S., UK, Singapore and Germany. The data from social media have also received a lot of attention in economics field, especially stock price prediction research. Stock price is dynamically changed, which is based on real-world events and people's point of view. Therefore, social media could be the most effective resource to predict the stock price trends. Sprenger considered twitter as a forum that leverages the wisdom of crowds for extracting the stock-related opinions. The correlation between stock market events and social media activities are applied to stock price prediction (Ruiz, Hristidis, Castillo, Gionis, and Jaimes, 2012). Sentimental analysis are also successful factor in stock market forecasting (Bollen, Mao, and Zeng, 2011). They examined how emotions can actually affect decision making in stock markets. The proposed approach gave 86.7% accuracy in daily prediction. The above studies proved that twitter data is very helpful to detect and predict events in several research areas.

As people noticed that user data from social media sites well represents the people's interests, social media sites started to present the most discussed and searched topics, called trending topics, based on the data from their services (Naaman, Becker, and Gravano, 2011). Those trending topics services have received a great amount of attention. For example, 'Twitter Trending Topics', real-time event detection service provided by Twitter, shows the most often mentioned or posted short phrases, words, and hash-tags (Becker, Naaman, and Gravano, 2011). However, they show only the topic term and its rank with no detailed explanation. Hence, many researchers applied various summarisation and extraction approaches aimed at revealing the exact meaning of trending topics.

Sharifi et al. applied a phrase reinforcement algorithm to summarise related tweets of Twitter Trending Topics (Sharifi, Hutton, and Kalita, 2010). Then, the author conducted evaluation for comparing hybrid TFIDF and phrase reinforcement in use of Trending topics summarising. Additionally, there is a experimental study conducted for comparing twitter summarisation algorithms (Inouye and Kalita, 2011). They found that simple frequency-

based techniques produce the best performance in tweets summarisation. Han and Chung (2012) applied simple Term Frequency approach for extracting the representative keywords to disambiguate the approach. They also proved that the most successful approach to reveal the exact meaning of trending topics is simple Term Frequency, which is evaluated by 20 postgraduate students in 2014. Some researchers examined classifying trending topics. There is a study that classifies trending topics into 18 general categories by labeling and applying machine-learning techniques (Lee, Palsetia, Narayanan, and Patwary, 2011). It aimed to classify trending topics by applying several proposed features and used SVM to check the accuracy (Zubiaga, Spina, Fresno, and Martinez, 2011).

Various types of topics detection and prediction research in social media have been conducted. Nikolov and Shah (2012) proposed a new algorithm for early detection of trending topics on Twitter. The performance achieved 95% accuracy. However, trending topics ranks and the prediction of rank changes has never been investigated before. Myers and Leskovec (2014) explored how the burst affects the diffusion of posting in social media and developed a prediction model that forecasts which information diffusion events will affect the bursts in network dynamics. Some researchers proposed cascade growth prediction using various features, including content, root, resharer, structural, and temporal features (Cheng, Adamic, Dow, Kleinberg, and Leskovec, 2014). Zhang and Pennacchiotti (2013) analysed and predicted e-commerce behaviours (a user's purchase behaviours) using only social media information (facebook user's profile information).

5.3 Temporal Modeling of Trending Topic Ranking Changes

The goal of this research is to predict the trend of trending topics rank change in the next hour. I propose a temporal modeling framework for predicting trending topics rank change using historical rank pattern data and machine learning techniques. The proposed temporal model can be described using equation 5.1:

$$FRC(T_x) = ML(PRP(T_x)) \tag{5.1}$$

In order to predict the next rank change FRC of a specific trending topic T_x, I used past rank pattern data (PRP) of the topic T_x. Then, machine learning techniques ML are applied for learning our model. Equation 5.2 describes the example of historical rank pattern PRP of a specific trending topic T_x at time t. It shows all historical rank patterns of a topic T_x in the specific period n. FRC represents the trends of the topic's ranking in the next hour. By comparing the current rank and the next-hour rank, the predicted rank change in the next

hour will be one of three classes: up, down, and unchanged. For example, if the next-hour rank r_{t+1} is higher than the current rank r_t, the FRC will be 'down'.

$$PRP(T_x) = [r_{t-n}, ..., r_{t-1}, r_t] \tag{5.2}$$

$$FRC(T_x) = \begin{cases} up, & \text{if } r_t - r_{t+1} > 0 \\ down, & \text{if } r_t - r_{t+1} < 0 \\ unchanged, & \text{if } r_t - r_{t+1} = 0 \end{cases} \tag{5.3}$$

There are two main issues when I use the historical ranking data for our model: missing ranking handling and window size selection. Firstly, several trending topics tend to disappear and reappear from the 'Trending Topics' list. We specify how to handle missing rank values during the topic's disappearance. Secondly, the historical ranking patterns of trending topics are time-series data so it is crucial to select the appropriate window size for prediction. We propose an approach to select the optimal window size of our data. The detailed information of these approaches for those issues can be found in the following sections: 1) Missing Ranking Handling and 2) Window Size Selection

5.3.1 Missing Ranking Handling

As the 'Trending Topics' list displays the top 10 trending topics of the moment, it displays the topics from rank1 to rank10. In other words, if the topic is suddenly out of the 'Trending Topic' list, it is impossible to recognize the exact ranking, whether the topic is ranked 11th or 50th.

Figure 5.4 and 5.5 shows the example of the nature of trending topics disappearance and reappearance. Those two figures represent the hourly rank change of two different trending topics in 24 hours; x-axis represents the 24 hours from the point the trending topic initially appeared on the list, and y-axis shows the ranking, from rank1 to rank under 10th, of the trending topic. 'Under 10' in y-axis describes when the topic disappeared from the list. Figure 5.4 shows the hourly rank change of the topic '#iPhone5s'. The topic appeared when Apple introduced the iPhone5. The topic was out of the list for three hours from the point it appeared in the 16th hour. Figure 5.5 shows the rank change of the topic 'Beyonce' that is referring to news that Beyonce had a fight with her husband. We revealed a common pattern that trending topics disappear and reappear on the list. When the topic reappeared again, it had a similar ranking to the point it disappeared.

Manual inspection of the trending topics revealed that topic disappearance-and-reappearance is not limited to the type of topic. Various types of trending topics, including breaking news,

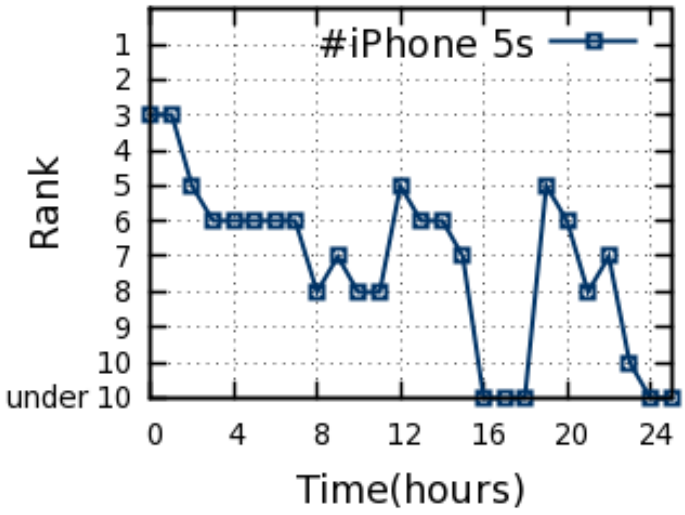

Fig. 5.4 Topic disappearance and reappearance pattern of Topic "#iPhone5s" from Trending Topics list

persistent news (e.g. TV show or sport match) and hash tags seem to disappear and reappear randomly. We then analyzed how many trending topics actually disappear and reappears. Table 1 shows the percentage of trending topics that reappear and failed to reappear after the topic disappeared. The proportion of reappearing trending topics is almost 70%.

Table 5.1 The percentage of trending topics that reappeared or non-reappeared after it disappeared

	Reappearance	No-reappearance
Pecentage	66.28%	34.82%

Based on this analysis, I claim that it is crucial to deal with missing rank data for our prediction model. We applied the following four missing value-handling approaches reviewed by Allison (2000):

1. Deletion
2. Dummy variable control
3. Mean substitution
4. Expectation maximization

The prediction results of these approaches will be discussed in the 'Evaluation Result'.

The following shows the detailed explanation of using four different types of missing value handling.

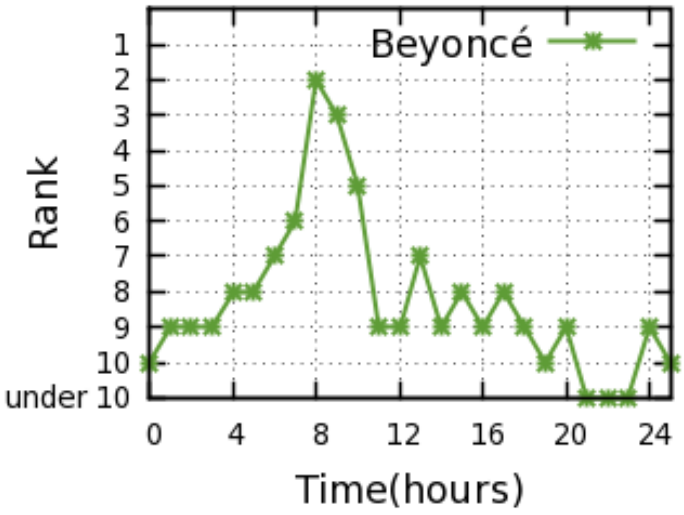

Fig. 5.5 Topic disappearance and reappearance pattern of Topic "Beyonce" from Trending Topics list

1. **Deletion**: There are two different types of deletion methods: listwise deletion and pairwise deletion. Listwise deletion is the method that removes entire records that contain any single missing value. Pairwise deletion analyses all cases in which the variable of interest is present, and uses all information possible for each analysis so it is more likely not to have a bias in estimation. Hence, I applied the pairwise deletion approach for handling the topic disappearance.

2. **Dummy variable control**: This approach sets up an indicator for missing value. It should impute the missing value to a constant. It can use all available information about missing observations but it is not theoretically driven. For our research, I replace the rank of the topic disappeared to zero (0).

3. **Mean substitution**: This is the method to replace all missing values in a variable by the mean of that variable. Using mean substitution is based on the fact that the mean is a reasonable guess of a value for a randomly selected observation in a normal distribution.

4. **Expectation maximization**: This is a maximum likelihood approach that can be used to create a new data set in which all missing values are imputed with a maximum likelihood value. The single imputation using EM identifies the value that produces the highest log-likelihood. Based on the EM calculation, I found that the replaceable value for the missing value of our data should be the rank, lowest+1.

5.3.2 Window Size Selection

The proposed temporal model uses historical trending topics and is learned using machine learning techniques, so it is important that sequences of the same window size should be used in training and testing. However, the primary difficulty is selecting an optimal window size for prediction using a good learning technique instead of trial and error.

We analyze the actual trending topic ranking data on USA Twitter. According to the data analysis result, I found that the same topic terms are sometimes referring to different events, and this normally occurs when the time length of the topic disappearance exceeds a certain time. For example, table 2 shows the example of analyzing the same trending topic '#MalaysiaAirlines' that is about two different events. The table displays the collected date and representative content of each topic. In 2014, there were two sad events that are related to Malaysia Airline: Firstly, the Malaysia airline flight MH370 disappeared on 8 March carrying 227 passengers and 12 crews. The second referred to MH17 which is believed to have been downed by a surface-to-air missile in the eastern Ukraine on 17 July with 259 passengers on board. The table shows that the same topic '#MalaysiaAirlines' are about two different events based on the collected date. On the first row, the extracted content shows the words missing, disappear and loss, which are related to the missing airline. The extracted contents on the second row has the words shot, missile, kill, crash, attack and explode which relates more to the airplane that was allegedly attacked by missile. Hence, before and after almost 4 months disappearance, the topic term '#MalaysiaAirlines' is separated into two different events.

We proposed the approach to identify the minimum length of topic disappearance that has different contexts by comparing the context similarity in two time-points (before-and-after the topic disappearance). As mentioned earlier, the trending topic terms consist of words, hash-tags or short phrases but it does not provide any description. It is almost impossible to recognize the exact meaning of a trending topic without extracting its detailed information. Hence, I proposed an approach to extract the representative contents for each trending topic and compare the context similarity in two time points (before-and-after the topic disappearance) if the topic disappeared at one point. The proposed approach is conducted as follows:

1. collect the trending topic and related tweets of the topic

2. preprocess the related tweets by removing stop words

3. extract the representative 15 (fifteen) terms using term frequency (TF)

4. 4) calculate the context similarity of a specific trending topic at two different time-points (before- and-after the topic disappearance).

Figure 5.6 and figure 5.7 that show the result of context similarity based on the length of continuous disappearance; x-axis represents the length of topic continuous disappearance, and y-axis shows the cosine similarity rate (1 means exactly same and 0 is completely different). Figure 5.6 displays the simliarity changes based on the time disappeance time (hour) and figure 5.7 shows the difference based on the time disappeance time (day). As you can see from the graphs, you can find that the context similarity is very low (0.2) if the topic continuously disappeared for over 7 hours.

Moreover, the similarity does not go down after 7 hours, which is around 0.2. In other words, if a specific trending topic 'A' does not appear in the list for over 7 hours and then reappears again, I can tell the first appeared topic 'A' and reappeared topic 'A' are talking about different contexts. In other words, if the topic disappears for less than 7 hours and reappears, the topic can be considered as the same topic. If the topic disappears for more than 7 hours, the topic is considered a different topic.

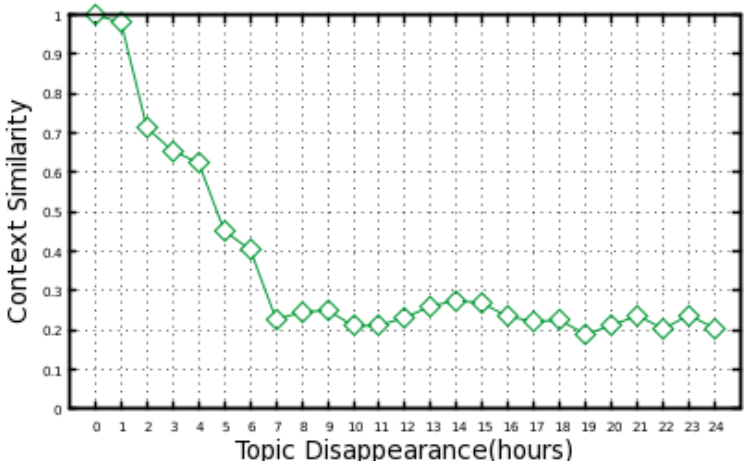

Fig. 5.6 The average of content similarity based on the topic disappearance time (hours)

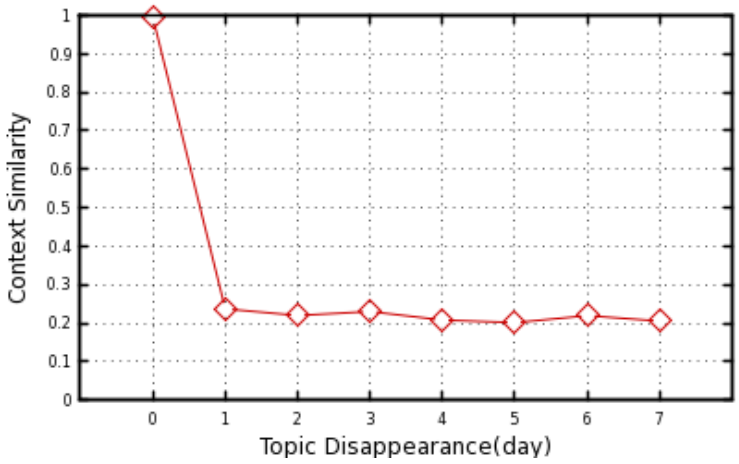

Fig. 5.7 The average of content similarity based on the topic disappearance time (days)

Based on this result, the optimal window size for trending topic rank data can be the minimum length of continuous disappearance without different contexts in same topic term. The optimal size for USA twitter trending topic rank data should be 7 (seven). The evaluation of prediction with different window size will be conducted in the evaluation result.

Table 5.2 U.S Trending Topics Ranking Change Prediction Accuracies with Different Missing Ranking Handling Approaches and Window Sizes

	# of instances	Missing Value	NB	NN	SVM	C4.5
(1)	5	Zero(0)	79.71%	88.20%	79.91%	88.74%
(2)	5	Lowest+1	80.11%	88.92%	80.82%	89.85%
(3)	5	Mean	75.10%	86.56%	77.29%	87.49%
(4)	5	Deletion	75.91%	85.42%	77.52%	85.74%
(5)	7	Zero(0)	83.91%	93.56%	85.36%	93.08%
(6)	**7**	**Lowest+1**	83.03%	**93.68%**	86.04%	**94.01%**
(7)	7	Mean	80.23%	91.06%	83.22%	92.91%
(8)	7	Deletion	82.93%	92.76%	83.93%	90.10%
(9)	9	Zero(0)	83.88%	92.53%	85.31%	93.00%
(10)	9	Lowest+1	83.00%	92.54%	85.61%	93.88%
(11)	9	Mean	80.34%	91.40%	83.29%	92.14%
(12)	9	Deletion	82.91%	90.92%	83.91%	90.11%

5.4 Experimental Set-up

I describe the collected data and applied machine learning techniques for evaluation of trending topics‘ rank change prediction. Algorithm 1 shows the whole algorithm for prediction. Based on Algorithm 1, I prepare the required data and machine learning techniques.

5.4.1 Evaluation Data

For the evaluation, I collected trending topic terms, related tweets and ranking patterns for those topics. By using the Twitter API, I crawled trending topics, related tweets and their ranks for two years (from 30th June, 2012 to 30th June, 2014) in different countries (USA, UK, and Australia).

Trending Topics

Twitter monitors all tweets, detects the top 10 most popular topics of the moment, and publishes those on the ‘Trending Topics’ list. It is located on the Twitter interface by default so all twitter users can check the current trending topics and discuss them. According to the previous studies (Kwak et al. 2010), trending topics represent the real-world breaking news. When the celebrities die or there is a major disaster, most trending topics reflect that news. For example, when Robin Williams died on August 12th 2014, most trending topics on that day were about his death and movies, such as RobinWilliams, Mrs.Doubtfire, Flubber, etc. For this research, I collected the top 10 trending topics every hour via Twitter search API. In total, I have collected 63404 unique trending topics over two years.

Algorithm 1 Trending topics popularity prediction

1: Collect a trending topic T from Twitter with the rank r and the collected time(hour) h.
2: Put a topic T using search API and obtain the related tweets rt that are posted at the time $h-1$ to h
3: Check whether the topic T appeared on the list before. If so, extract the representative words that describes the meaning of collected trending topics using Term Frequency. If not, the topic T is the new topic so skip to step 5.
4: Obtain all previous ranks PR of the trending topic T from the time(hour) $h-n+1$ to h (n=window size).
5: Use this previous ranks PR as input data to the models trained by machine learning techniques
6: Predict the rank change FRC of the trending topic T will be up, down, or unchanged in the next hour.

Related Tweets

Since the trending topics list displays only the topics terms, with no detailed information, it is almost impossible to identify the meaning of trending topic until you examine related tweets for that topic. For example, when a missile allegedly destroyed Malaysian Airlines on July 17th, 2014, the topic '#MalaysiaAirlines' appeared on the list. Since it provides only the topic term without any explanation, users may realize that something happened to a Malaysian Airlines flight but what exactly happened is not clear. To reveal the exact meaning of the trending topics '#MalaysiaAirlines', users need to read all related tweets for that topic, which is impossible. For disambiguating the exact meaning of trending topics automatically, I searched the related tweets of each trending topic. While collecting the related tweets, I aimed to avoid crawling the tweets that contain irrelevant contents. If the trending topic is '#MalaysiaAirlines' which relates to the Missile attack on July 17th, I should not collect any tweets about 'Missing Malaysia Airline' occurred on March 8th. Due to this issue, it is extremely important to distinguish the related tweets. Twitter API provides the tweet/search crawling service that allows users to collect the tweets with detailed information of each tweet, including content, username, location, created date-time, etc. We used this created date-time to extract the appropriate tweets for the trending topics. As I collect the top 10 trending topics on an hourly basis, I search and collect the related tweets that users upload in the last one hour. For example, when 'Malaysia Airline' is on the trending topics list at 8pm, I search and collect the related tweets that users uploaded between 7pm and 8pm. This collecting approach minimizes irrelevant tweets.

5.4.2 Machine Learning Techniques

From the previous sections, I mentioned how to collect the required data. We applied the machine learning techniques for building the prediction model using our data. Machine learning techniques are initially introduced for predictions based on known properties learned from the past data.

In this study, I used the historical ranking pattern data of trending topics for prediction, so machine learning techniques are very suitable for this application. We selected four machine learning approaches: Naive Bayes, Neural Networks, Support Vector Machines and Decision Trees. The philosophies behind these four algorithms are very different, but each has been shown to be effective in several time-series prediction studies.

Naive Bayes

In machine learning, naive bayes is one of the successful approaches for classification. It uses probabilistics for predicting the result by analysing cause-and-effect relationship from the past data. The approach is very effective when the cause is not recognisable. Naive bayes approach is based on the Bayes' Theorem (Efron, 2013), results of improvement of conditional probabilistics by analysing more and more new information with strong independence assumptions. The Bayes' theorm can be written as:

$$p(C|F_1,...,F_n) = \frac{p(C)p(F_1,...,F_n|C)}{p(F_1)} \tag{5.4}$$

over a dependent class variable C with several feature variables F1 to Fn. The above equation represent the posterier probability calculation, prior probability multiplied by likelihood is divided by evidence. Naive Bayes is generally used for stock prediction and sales prediction studies (Sprenger, Tumasjan, and Sandner, 2013). In this research, I applied Naive Bayes classifier with estimator classes. Numeric estimator precision values are chosen by analysing the training data set, ranking patterns.

Neural Networks

Neural Networks (NN) approach, often called Artificial Neural Networks (ANN) approach, is a learning algorithm that is derived from the structure and function of biological neural networks. NN is considered as very successful technique in classification, prediction, and pattern detection studies. NN consists of interconnected group of nodes, named after 'Neurons' in the brain. As can be seen in figure 5.8, it has three layers, including input, hidden, and output layer. The first layer has neurons for sending the data via synapses to the hidden layer, and then via more synapses to the output neurons. In complicated systems, they have more layers of neurons, which contain increased layers of input neurons and output neurons. While the calculations, the data will be manipulated by the weights, which are synapses store parameters (Arauio, Astry, FerrerioLage, Mejuto, RodriguezSuarez, and Soto, 2011). A neural networks are defined by three types of parameters:

1. the interconnection between the different layers
2. the learning process for updating the its weights
3. the activation function, which transforms a neuron's weighted input to the output.

We applied multilayer perceptron(MLP), also known as feed-forward artificial neural network model (FFNN), which consists of multiple layers of nodes in a directed graph, with

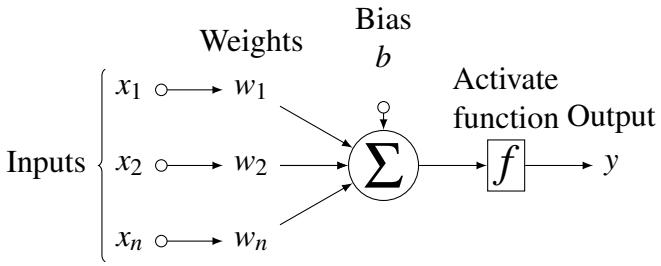

Fig. 5.8 Neural Networks

each layer fully connected to the next one. FFNN uses a backpropagation for training the network. This approach is the advanced version of the traditional linear perceptron and can distinguish non-linear data. It is considered one of the most successful NN techniques that provides the highest performance.

Support Vector Machine

Support Vector Machines(SVM) is a group of supervised learning algorithms that is used for several research areas: data analysis, classification, regression, and pattern recognition (Abe, 2010). As you can see the below figure 5.9, the idea of SVM is finding optimal hyperplane for linearly separable patterns from the given training data. Unlike other machine learning algorithms, including Neural Network, SVM does not finds hyperplane that just separate the data points but also has maximum-margin.

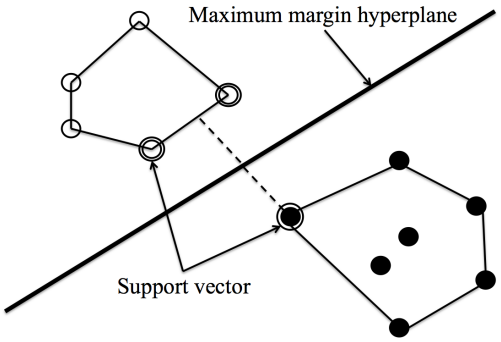

Fig. 5.9 Support Vector Machine

The margin is the minimum value of distance from the hyperplane to each data point. In order to classify data points into two different classes with the maximum margin, the hyperplane should be located at the point, which has the same minimum distance from both the two different classes. Therefore, SVM can be a model that represents the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. As well as performing linear classification, it can be used

for non-linear classification by applying kernel trick(Abe, 2010). In pattern matching, SVM generally performs better than neural networks or decision tree learning algorithms. This is because SVM aims to achieve not only the highest classification performance but also the maximum margin. In this research, I used Linear SVM.

Decision Trees Learning

Decision Tree is a tool that supports decision-making by using a tree-structured model of decisions. It can be separated into two types of trees: classification tree and regression tree (Loh, 2014). Classification tree is for predicting the class outcome which the data belongs, and regression tree is when the predicted outcome is the numerical value. The tree is usually made by recursive partitioning approach, and can be read as a flowchart that has four different types of elements: root node, internal nodes, link, and leaf: each node contain the criteria for classification, and each leaf node shows a class. The link from root to leaf node represents classification rules that lead to the classes. It maps the features by using characters or numbers, which represents a class. The tree does not output the decision but offers support by representing the data. Decision tree is generally used in decision-support for classifying items or evaluating processes. Decision tree learning approach utilises the decision tree as a predictive model in machine learning, data mining and statistics. It constructs a decision tree from the class-labeled training data set for predicting the value of a target variable based on several input data.

There are various decision-tree generation algorithms, including ID3, C4.5, MARS, etc. For this study, I applied C4.5 decision tree generation algorithm, which is an extended version of ID3 algorithm. C4.5 algorithm generates the decision trees for classification so the algorithm is often considered as statistical classifier.

5.5 Evaluation Results

In this section, I now compare and summarize the prediction results with the proposed temporal model and different missing ranking handling approaches and window sizes.

5.5.1 Window Size Selection Examination

As I discussed in the methodology, I proposed that the approach to selecting the optimal window size for trending topics' ranking change predictions. We found that optimal window size can be same as the minimum length of topic disappearance time that has same topic

term with different meaning (refer window size selection in methdology). We discovered the optimal window size for U.S twitter data can be 7(seven).

Table 5.3 Optimal window size for three countries (United States, United Kingdom, and Australia)

	USA	UK	AU
Optimal Window Size	7	6	8

In order to examine the proposed window size selection approach, I applied our approach to the trending topic rank data from U.K. and Australia Twitter. Based on this examination, I found that the optimal window sizes for U.K. and Australia was 6(six) and 8(eight) respectively, as shown in Table 5.3. We evaluate the prediction performance with those window sizes to examine whether the proposed approach selects the optimal window size of different data.

5.5.2 Prediction Evaluation

The experiments were designed to test the proposed model. We use the prediction performance as an indication of the suitability, which is obtained from four machine-learning techniques I discussed in the previous section.

Each experiment result has different window sizes and different missing ranking handling techniques. Table 3 shows the prediction result of U.S. trending topics ranking changes with different window sizes (5,7,9) and four different missing handling techniques (Zero, Lowest+1, Mean, and Deletion). As mentioned in 'window size selection' section, I insist that the optimal window size for USA trending topics rank data can be size 7(seven). The experiment result shows that the prediction with size 7(seven) has the highest performance among 5, 7 and 9, which proves that our approach performs successfully. Since there is little difference in prediction accuracy of size 7 and 9, it is difficult to define whether 7 is better than 9. However, I can infer that if there is no difference between size 7 and 9, using size 7 is effective in performance, including data size and speed.

For missing ranking handling, single imputation techniques, especially lowest+1, show better performance than the others. As you can see all three instances (5,7,9), missing value imputation with lowest+1 achieve the best prediction performance. This is because the other three approaches, mean, zero, and deletion, are not considered the nature of trending topics ranking but the single imputation with EM, lowest rank+1, analysed the trending topic rank to extract the appropriate missing value. Therefore, it shows the best prediction accuracy in all three instances.

We applied four machine learning techniques, including Naive Bayes, Neural Network, Support Vector Machine, and C4.5 Decision Tree algorithm. Among 4 approaches, C.4 algorithm showed an extremely higher performance than the others. Finally, I analyzed the performance of two more countries (U.K. and Australia) to make sure that our model performs well. In the previous section, I found that the optimal window sizes for two countries were size 6 and size 8 respectively. Based on the each experiment result, U.K. Trending Topic Ranking change prediction achieves the best prediction performance (92.54%) with 6 instances and lowest+1 imputation, and Australian results the highest accuracy (80.13%) in 8 instances and lowest+1 imputation, which are exactly same as what I expected.

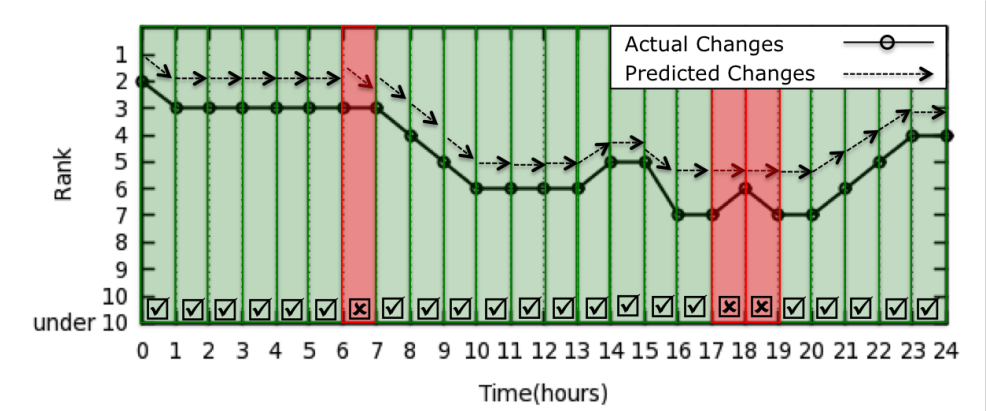

Fig. 5.10 Trending Topics 'Noel Pearson' Rank Change Prediction

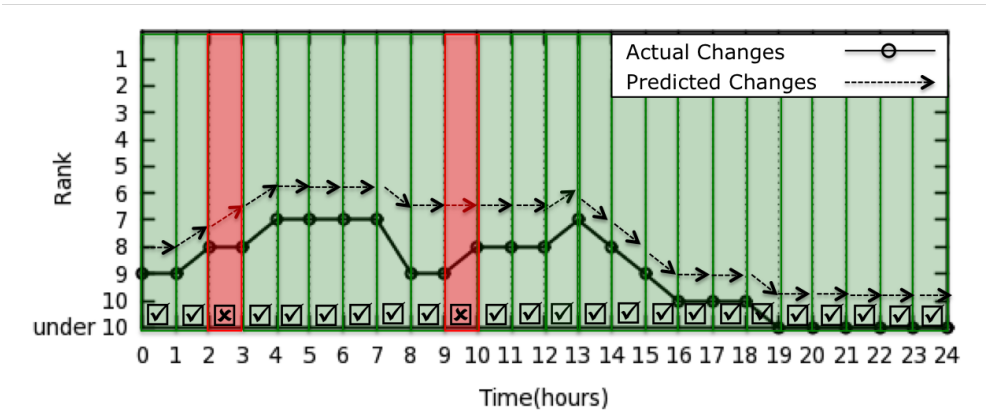

Fig. 5.11 Trending Topics 'Jamies Winston' Rank Change Prediction

Figure 5.10 and figure 5.11 presented that the examples of rank change prediction in 24 hours, which start from the point the topic first appeared on the trending topic list. As I mentioned before, our prediction model forecasts the next-hour rank changes of trending topics using only historical data with proposed missing value and window size treatment

approach. Both figure 5.10 and 5.11 show that the proposed model successfully predicted of rank changes. The prediction model works well in most cases except the noise problem. The noise is the pattern just temporal or irregular rank changes but does not affect to the broad rank changes. For example, the figure 5.10 has very short noise in the pattern from hour 17 to 19, and there are two noises, hour 3 to 4 and hour 9 to 10 in the figure 5.11.

5.5.3 Additional feature

We put the additional features (topic features of the trending topic) into the training dataset, and compared the accuracy with that of the historical rank data. We further classified the U.S. trending topics using the New York Times (NY times) classification service. As Trending topics are about real-time events, the traditional topic classification ontology cannot be applied. Unfortunately, if the category of a trending topic is extracted using a general document ontology, at any time, semantically related categories will also be extracted. The way I identified the category of a trending topic is as follows: first of all, I search the trending topic's term with the NY times topic classification service. We set the published time as the day that trending topic first emerged. Then I can locate any related articles that were published with that term, on that day. Finally, the trending topics related categories are supplied by the NY Times classification service. Table 5.4 shows how U.S. trending topics are categorised.

Table 5.4 Topic disbribution in U.S. Trending Topics

No	Topic	Percentage
1	Entertainment	42%
2	Sports	28%
3	Politics	10%
4	Fashion	6%
5	World issue	5%
6	Obituaries	4%
7	Health	3%
8	Technology	2%

After I added this topic attribute into the training dataset, I learned the model with C4.5 decision tree algorithm. The accuracy with topic attribute was 94.85%, which is slightly higher than that with historical rank pattern only (94.01%).

5.6 Implementation

The following figure 5.12 shows the architecture of the research.

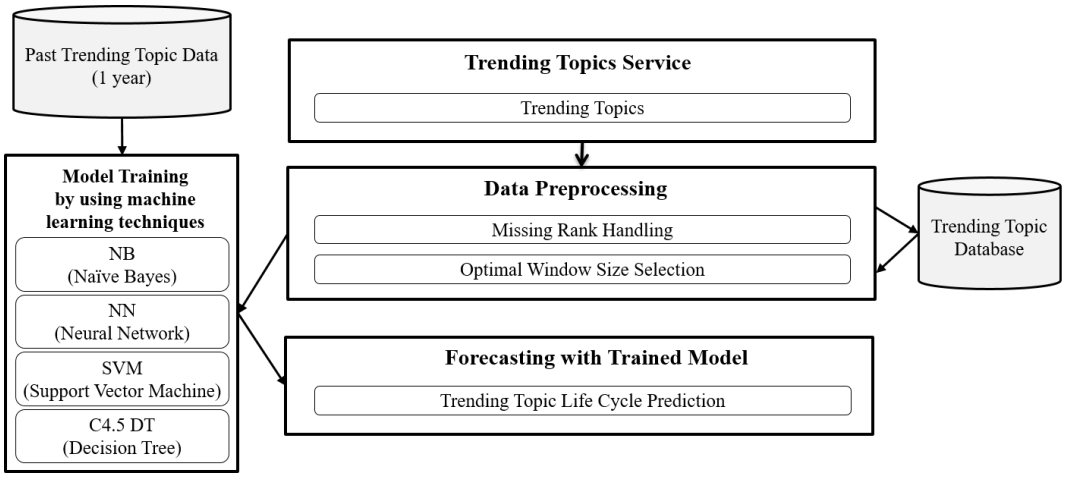

Fig. 5.12 System Architecture of trending topic lifecycle prediction

The database for this research is designed for storing all detailed information of trending topics from Twitter trending topics analytics services with the related information. Meaning of each stored trending topic is disambiguated by using related information and store in the database. The rest of tables are for storing machine learning prediction accracy evaluation, which contains the detailed information of window size selection and missing ranking handling approach. The database for this project is designed as follows:

- Table: tb_twt_keyword
  - id: primary key, the identification number generated by auto increment function.
  - keyword: twitter trending keyword
  - rank: rank of the twitter trending keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - country: country of the twitter trending keyword
  - local_time: local time at collection
  - date: collect time
- Table: tb_twt_relatedTweets
  - id: primary key, the identification number generated by auto increment function.

  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - tweet_id: the identification number of the related tweet given from twitter api
  - tweet_content: contents of the related tweet
  - tweet_date: uploaded time of the related tweet
  - retweet_count: the number of re-tweet of the related tweet given from twitter api
  - favorite_count: the number of favorite of the related tweet given from twitter api
  - date: collect time
  - tb_twt_relatedTweet_user_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweet_user table

- Table: tb_twt_relatedTweet_user
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_relatedTweet_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweets table
  - user_id: the identification number of the twitter user given from twitter api
  - user_name: the identification username of the twitter user
  - user_screenName: the screenname of the twitter user
  - user_location: the location of the twitter user uploaded tweet
  - user_followers_count: the number of followers of the twitter user given from twitter api
  - user_friends_count: the number of friends of the twitter user given from twitter api
  - date: collect time

- Table: tb_twt_relatedNews
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - news_content: contents of the related news
  - news_date: uploaded time of the related news

  - source: the source of the collected related news
  - date: collect time

- Table: tb_keyword_meaning_disambiguation
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - keyword: twitter trending keyword
  - content_tweet_kfe: key factor extraction result of the related tweets
  - content_tweet_ner: named entity reconiser result of the related tweets
  - content_tweet_tm: topic modeling result of the related tweets
  - content_tweet_as: automatic summarisation result of the related tweets
  - content_news_kfe: key factor extraction result of the related news
  - content_news_ner: named entity reconiser result of the related news
  - content_news_tm: topic modeling result of the related news
  - content_news_as: automatic summarisation result of the related news
  - content_combined_kfe: key factor extraction result of the related news and tweets
  - content_combined_ner: named entity reconiser result of the related news and tweets
  - content_combined_tm: topic modeling result of the related news and tweets
  - content_combined_as: automatic summarisation result of the related news and tweets
  - date: collect time

- Table: tb_machine_learning_evaluation
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - approach: machine learning approach
  - window_size: window size of the rank patterns
  - missing_handling_type: missing ranking handling type of the rank patterns
  - accuracy: prediction accuracy of the approach

  - date: collect time

- Table: tb_window_size_results
  - id: primary key, the identification number generated by auto increment function.
  - keyword: twitter trending keyword
  - unique_pattern_id: the identification number of each separated set of rank patterns
  - window_size: window size of the rank patterns
  - similarity: similarity of the content

- Table: tb_missing_ranking_handling_results
  - id: primary key, the identification number generated by auto increment function.
  - keyword: twitter trending keyword
  - missing_handling_type: missing_handling_type of the rank patterns
  - accuracy: prediction accuracy

5.7 Application

Based on the evaluation result, I developed a trending topics popularity trends prediction system, called TrendsForecast[1], which represents the rank changes of trending topics in 10 different countries, U.S., UK, Australia, Canada, New Zealand, Philippines, Japan, Malaysia, Singapore and Korea. The proposed system, TrendsForecast, applied C4.5 model with the best window size and missing value imputation method for each county.

The table in the system, can be seen in the figure 5.13, presents not only future rank changes, but also various factors, including the time the trending topic first appeared, past rank changes in last 3 hours and a hour. Therefore, users can see the popularity and importance of a specific trending topic by observing its historical rank changes, as well as its future rank changes. The system also provides the 'upcoming' trending topics, which keywords have disappeared but have a chance to reappear in the trending topics list.

Figure 5.14 displays the historical rank changes of top 10 trending topics in last 25 hours. X value indicates the time, and Y value indicates the ranking of each trending topics keywords. It enables users to see how the popularity ranks were changed, as well as the time that topics are disappeared and reappeared. Users are able to download and print the historical rank change pattern of top 10 trending topics.

[1]TrendsForecast 2014 https://www.trendsforecast.net

Rank	Keyword	Past			Forecast	
	Current	Appearance (ago)	Changes (3 hours)	Changes (1 hour)	Remaining time	Trend (1 hour)
1	#FOURHANGOUT	7 hr(s)	–	–	7 hr(s)	–
2	Wayne Goss	3 hr(s)	–	–	2 hr(s)	↓1
3	#congrats5sos	2 hr(s)	–	–	3 hr(s)	↑1
4	#CamsNewVideo	7 hr(s)	–	–	8 hr(s)	–
5	Green and Gold	0 hr(s)	new	↑3	4 hr(s)	–
6	Richie Benaud	1 hr(s)	↑1	↓1	3 hr(s)	↓2
7	#vote5sos	3 hr(s)	new	↓1	2 hr(s)	–
8	#MyFOURQuestion	4 hr(s)	new	↓1	2 hr(s)	↑2
9	Aaron Rodgers	0 hr(s)	new	reappear	1 hr(s)	–
10	FIFA	1 hr(s)	↓2	↓1	1 hr(s)	disappear

Upcoming : chicago weather / daytona 500 / jk rowling / mike wallace / danica patrick / the national enquirer / j.k. rowling / lent / kate gosselin / match play

Fig. 5.13 Screenshot of TrendsForecast, Trending Topics Rank Change Prediction System

5.8 Discussion

In this chapter, I addressed trending topic rank prediction problems. The only available data for this problem is rank history data of each trending keywords. Therefore, people may have question about whether any predication models using this data can suggest any promising prediction results. This paper suggests a simple rank prediction that uses historical data with consideration of window size and missing value treatment. Surprisingly, our method achieved very significant performance (about 94% accuracy with C4.5 decision tree). On the one hand, this implicitly implies that the changing trends are the most important factors for rank prediction. On the other hand, it would be possible to improve performance of rank prediction. For example, it is possible to obtain additional information by analysing the related tweets retrieved by querying the trending topics to Twitter or by using the linked content within the related tweets. However, it would be very difficult to predict rank perfectly (100% accuracy), which is not because of algorithmic factors but because of trending topics' irregularly changing nature.

5.9 Conclusion

In this chapter, I proposed a temporal modeling framework that predicts trending topics' hourly ranking change. We developed the learning procedure that can be used to construct

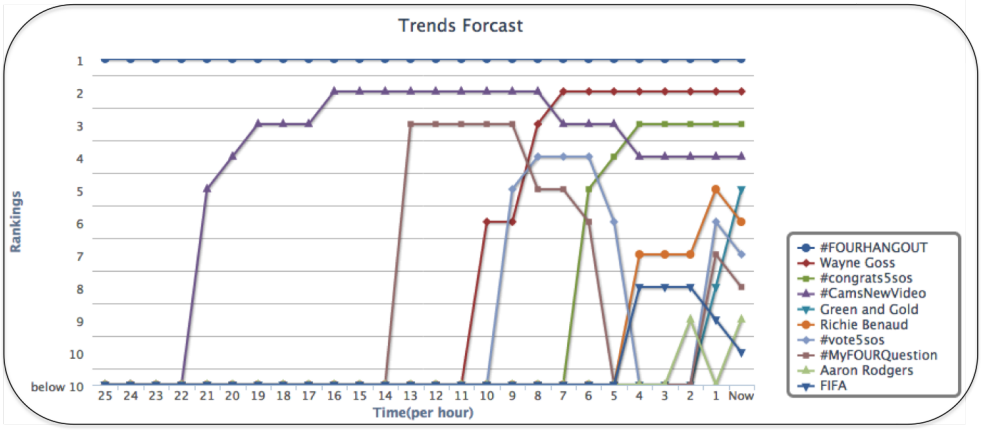

Fig. 5.14 Screenshot of TrendsForecast,Lifecycle Summary of Current Trending Topics

models of trending topics ranking trends based on the historical trending topics ranking patterns. We also suggested the novel approaches to handle missing ranking and window size. Rather than using complex features, I used historical ranking pattern and machine learning techniques, and it achieved the successful result (94.01%). The research is the initial work that proposes the temporal model of predicting of trending topics ranking as a degree of people's interests, and it achieves the successful result.

Chapter 6

Trending Topics Diffusion Prediction Among Services

Chapter 7 proposes diffusion predicting model for forecasting how trending topics diffuses among three different online web services, including search engine, social media, and Internet news service.

6.1 Introduction

In recent years, people share, search and obtain information by using different types of webbased services, such as search engine, social media, and Internet news aggregation sites. Those web services caused a huge informationsharing paradigm shift by increasing personal information sharing and obtaining. This phenomenon, called 'social data revolution', has resulted in the accumulation of unprecedented amounts of social data. This large amount of personal and social data, which is made by users themselves, is like a vein of gold in 21th century. Many internetbased companies realized the importance of analyzing this huge amount of social data since it is directly related to the users' interests and communications of their web services. Based on those data, many companies provide new services, which display the overview of the topics and issues that are currently popular within their own community. I call this service as trending topic analytic service (TTservice) since it represents the popular topics of communication among users on the services. In recent years, various internetbased companies, such as search engines, social networking services, and Internet news, provide TTservice. For example, Twitter provides the trending topic analytic service, called Twitter Trending Topics, which displays the list of the top 10 fastest rising discussed terms on Twitter.

Hence, by using the services, I can notice the issue that people concern within the certain online community, which can be called as 'Trending topic'. Unfortunately, most SASs currently provides only the emerging issue terms or keywords only in their community, not compare or integrate with those in other online communities. Therefore, it is hard to find how the trending topic diffuses from services to service on the web. Then, "is there any method to find out how those trending topics diffuse from web service to web service?" The question represents the motivation of this research. The research question of this chapter is "Can I detect and predict the trending topic diffusion process on the web?" More specially, I researches what types of trending topic diffusion are occurred on the web? What kinds of trending topic the users are interested in each services? Finally, how can I detect and predict the trending topic diffusion process from webservice to web service?

To solve the questions, in this research, I investigate the realtime trending topic diffusion among different web services. First of all, I collected trending topic keywords from different web services (search engine, social media, and Internet news aggregation sites) for one year. Then, I characterized the trending topic diffusion process based on the various features that are related to the information diffusion process. To do so, I researched on finding the type of features or factors, which are influence for the trending topic diffusion process on the web. As trending topic diffusion on the web has never been considered before, I find the appropriate features from the general information diffusion research field.

There are various kinds of features, which influences to the Information diffusion process on the web. SeedProvider of information also affected to the way of information propagation [17]. Topic feature has been proved that it is very affected to the information diffusion (Romero, Brendan, and Kleinberg, 2011). Hence, the information with different meaning and topic shows different diffusion process. Information diffusion is influenced by the starting time [6] as well. Based on these strategies, I characterize and model the trending topic diffusion process with those three features:

- Provider: Seed Provider. Starting point of trending topic. It represents the service that the trending topic was first emerged. I identify three types of web services that deal with the real time information, including search engine, social media, and Internet news aggregation site.

- Time: The feature defines the time that the trending topic started. The time is classified into three categories, including Night Time (1am to 9am), Business Time (9am to 5pm), and After Work Time (5pm to 1am).

- Topic: This feature represents the topic of the trending topic keyword. I first classify the topic of each trending topic keyword based on the meaning of the issue.

Subsequently, I validate our features on trending topic diffusion prediction model, which applied 8 types of machine learning techniques, including Fuzzy Unordered Rule Induction Algorithm (FURIA), Support Vector Machine (SVM), Knearest neighbour (KNN), C4.5 Decision Tree (C4.5), Ripple Down Rules (RDR), Kstar, Feed Forward Neural Network (FFNN), Logistic Regression (LR). The contribution of the research are summarized as follows:

- The research provides a characterization of trending topic diffusion based on topic, time, and service.
- The research is the first ever study on the trending topic diffusion model through the web services. Trending topic diffusion model has never been reported before in the literatures.

6.2 Related Work

6.2.1 Information Propagation

Nowadays, social media replaces the status of blog world such as microblogs which are easier to manage with short comments. So the diffusion occurred not only in blog world, but also in social media. Due to the two different structures, the information diffusion process should be different. In hence, the data collection and data analysis for blog and social media should be diverse. The process of modelling information diffusion also will alter. On one hand, as I mentioned before, Kim et al. (2013) also thought that major diffusion models are derived from Bass model (Bass diffusion model 2014). They considered the influence of networks in social media into direct influence and indirect influence by two features and heterogeneity of populations, and generated Bass diffusion for the dynamics of meta-populations by conducting direct influence model and indirect influence model. On other hand, Bourigault et al. (2014) thought that the diffusion from the source to all directions in latent space, and proposed Diffusion Kernel that capture the dynamics of diffusion of the cascades. They named their model as 'content diffusion Kernel' (CDK). And they proposed use a classical stochastic gradient descent method, which iterates until a stop criterion is met. They thought that the different content will diffuse differently in network. In hence, they proposed to consider the content of each cascade into their model. Li et al. (2013) pointed out that GT model perform better than theory-centric models and data-centric models and proposed a social influence representation method to predict the temporal dynamics accurately. They also proposed time- dependent user payoff calculation method to calculate

the payoff of user facing its neighbour both global influence and social influence. The model presented by Bourigault et al. (2014) is learned by observation called Content diffusion Kernel like modelling of Leskovec et al. (2007) and McGlohon et al. (2007). Due to their CDK model is relying on the size of the latent space, but not based on any knowledge of structure of network, the larger space will predict in high quality as well as IC model. Nowadays, with the popularity of social media, researchers paid more and more attention of information diffusion on social media area. In order to study the diffusion in social media area, the major researchers choose Twitter as the source (Yang & Counts 2010; Romero et al. 2011; Kim et al. 2014; De Choudhury et al. 2010; Remy et al. 2013; Kwon & Han 2013; Taxidou 2013) to analyse data and experiment.

Table 6.1 Summary table for Information Propagation Research

Purpose	Data	Approach	Applied research
Discover mechanisms of information diffusion across different types of social networks	ICWSM 2011 dataset	Two macro-level diffusion models; direct influence; indirect influence; bass model	Kim et al. 2013
learn a mapping of the observed temporal dynamic onto a continuous space	International AAAI conference on Weblogs and Social Media 2009; Memetracker corpus; Digg	Content diffusion Kernel; IC model	Bourigault et al. 2014
Model the process of information diffusion in social networks	Sina Weibo; Flickr datasets	GT model; time-dependent user payoff calculation method; new social influence representation	Li et al. 2013
Predict the speed, scale, and range of information diffusion in Twitter	one month Tweets through the Twitter API	Observation	Yang & Counts 2010
Analysing the ways of hashtag spread in Twitter	Tweets of user with hashtags	Ordinal time; snapshot	Romero et al. 2011
Find influential neighbours to maximize information diffusion in Twitter	Twitter dataset related to the 2010 UK general election	Independent Cascade Model	Kim et al. 2014
Demonstrate all users are not equal on the aspect of information diffusion	Twitter dataset collected during the great Tohoku earthquake in Japan	Observation	Remy et al. 2013
Examine the factors behind social transmission	Twitter data with Korean users.	Poisson regression model	Kwon & Han 2013
Investigate real-time analysis methods on social media	Tweets from Olympics 2012 in London, US lections 2012 and Super bowl 2013	State-of-the-art lacks specific algorithms	Taxidou 2013

To investigate the information diffusion, Taxidou (2013) discerned that observing the information flow such as information cascades can predict the aspects of information diffusion. In hence, he collected over 800.000 retweets about US elections 2012 to generate a graph of cascades of the tweets recording Olympics. Additional, Romero et al. (2011) realized that the hashtags of twitters in Twitter are also an approach to collection data. In hence, they identified the 500 most mentioned hashtags at beginning, and then categorized these hashtags into eight categories at least 20 samples for one category. These categories of hashtags used for generating the curves to find out the stickiness and persistence of hashtags.

There is interesting theory provided by Kim et al. (2014) that a node is diffusing likely due to the active neighbours. Therefore, they tested the schemes from four selections containing random selection, degree selection, propagation-weight selection and hybrid selection. The user propagation weight, as a critical variable, means the average rate of each node propagating its neighbours. And the hybrid selection is the combined selection of other three, which results the most maximizing information diffusion (Kim et al. 2014).

Kim et al. (2014) were aware of the limitation while major researchers used the Independent Cascade model and the Linear Threshold model (Kempe et. al 2003) to infer the information diffusion. The IC model requires an initially activated node to infect other nodes, which results the researchers have to select a set of arbitrary nodes for initiation. However, epidemiological model proposed by Kim et al. (2014) is not based on the assumption of initial activated nodes. The nodes under this model are only interacting with their neighbours. They extended the model with several parameters containing user propagation weight, decay factor and content interestingness into the work of Kim and Yoneki who introduced the optimization problem to find influential neighbours for maximizing information diffusion.

6.3 Data Collection

I crawled and collected trending topics from three different services: search engine, social media, news aggregation website. The selected services are Google Trends, Twitter Trending Topics, and Google News Top Stories.

First, I collected search-engine based trending topics from Google Trends Hot Search, which displays the most popular search term for the past hour in the United States. Google Trends provides the additional information of each particular popular search term, which includes its last 24 hours search volumn graph, and the search result (blog, news, and webpage search) with the term. Social media services based trending topics are collected from Twitter Trending Topics, which represents the most discussed and posted topics in the service. However, Twitter Trending Topics do not provide any detailed information of topics

that shows the meaning of the trending topics. Twitter provides the search API that allows users to search the related tweet posting on a given query. News service based trending topics are collected from Google News, which is an one of the popular news aggregation service. It collects almost all news articles from different publisher in real time, and provides them. Google news do not monitor what the users are searching or discussing but analyses clicking behaviour. Based on the result of users' clicking activities, the service retrieves the most emerging articles, extracts the top ten nouns from the articles, and provides the 'Top Stories' service. I collected three datasets of trending topics from the three services for about one and half months (4 June, 2012 3 June, 2013) at one-hour interval. Table 6.2 summaries trend word collections. A total of 262,800 keywords there collected, but a total of unique words are as follows: Google Trends – 5162, Twitter – 20851, and Google News - 2350.

Table 6.2 Trending Topics Collections

Provider	# of trending topic terms	# of trending topic terms per day	# of distinct trending topic terms
Google Trends	87,600	14	5,162
Google News	87,600	57	20,851
Twitter	87,600	7	2,350

The table 6.2 shows that the number of distinct trending topics per day as follows: Google Trends – 14, Twitter – 57, and Google News – 7. Compare the freshness of trending topics among three services, Twitter updates the trending topic more frequently than other two services.

6.4 Methodology

I find the importance and usefulness of various features that are influenced to trending topic diffusion process. By characterizing and modeling the trending topic diffusion based on the three different features, including provider, time and topic.

- Feature 1 (Provider): The feature describes the service that a trending topic is firstly emerged. As mentioned before, I collected trending topic keywords from three different webservices, search engine (Google Trends), social media (Twitter), and Internet news aggregation site (Google News).

- Feature 2 (Time Class): The feature defines the time that the trending topic started. The time is classified into three categories, including Night Time (1am to 9am), Business Time (9am to 5pm), and After Work Time (5pm to 1am).

- Feature 3 (Topic): The feature indicates the topic of the trending topic. I have 9 topic categories to classify the trending topic keywords.

Based on this analysis, I used the above features to do:

1. Flow Prediction: the trending topic diffusion process from webservice to webservice

2. Time Interval Prediction: the time for trending topic diffusion process (how long does it take to diffuse from service to service?

6.4.1 Characterizing of Service

As mentioned before, I have collected and analyzed the trending topic keywords from three different webservices, which deal with realtime information, including search engine, social media, and Internet news aggregation site. Then, I become curious to know whether the trending topic diffusion process is affected the 'seed service', which the trending topic firstly received attention. For example, if a certain trending topic firstly emerged in the social media, would it show different trending topic diffusion process from that emerged in another service? To answer this question, I have identified each characteristics of service and examined whether the diffusion process is affected by the seed service.

Service Characterisation: I explore the characteristic of service in order to solve the above question. In this research, I cover trending topic from the top 10 popular trending topic terms in three different web services, including search engine, social media, and Internet news aggregation site. To understand the nature of the trending topic diffusion in different web services, I conducted analysis to find out the type of trending topic keywords are shown in the top 10 trending topic list and how the trending topics manifest and distribute in those three services.

Table 6.3 Trending Topics Terms Distribution

	Google Trends	Twitter	Google News
# of new trending topics	252	1184	617
# of common trending topics	1281	1867	1432
# of distinct trending topics	2350	20851	5162

Characterisation of Service based Diffusion

In order to use service factor as a feature for the service based diffusion, it is time to characterize the trending topic diffusion process based on the services.

To do so, I firstly integrated all the trending topic keywords from three services, Google Trends (search engine), Twitter (social media), Google News (Internet news aggregation site). Then, I extracted the common trending topic keywords among three services. The number of common trending topic keywords among three services is 2053 (out of 26000 – total number of trending topic keyword). Almost 10% of trending topic keywords are diffused among services.

Table 6.3 summarizes common trending topic keywords among three services. There are three rows to show different kinds of number of trending topic keywords. On the first row, you can see the number of new trending topic keywords, which describes the number of trending topic that starts from the Google Trends (the value of the first rowfirst column). The second row shows the number of common trending topic keywords, which means the number of trending topic keywords that have diffused to at least two services. On the third row of table 6.3 , it displays the number of distinct trending topic keywords that was in the top 10 trending topic list, which counted both common and uncommon trending topic keywords. From the table 6.3, I can see that the trending topic keywords from twitter are much fresher than other services, Google Trends and Google News. However, Google Trends provides very longlasting trending topic and tends to display the trending topics, which are already emerged from other services. It means people tend to search certain trending topic keyword that they have read from social media or news site.

I then become curious to know whether the startingservice are affected to the trending topic diffusion process. As can be seen from figure 6.1 to 6.6, I analyzed the diffusion process difference among different starting provider. From figure 6.1 to 6.3, it shows the percentage of trending topic diffusion, including three services diffusion process (over two nodes, e.g. GT TW GN), however, figure 6.4 to 6.6 displays the rate of two way trending topic diffusion process. For example, if the trending topic is firstly emerged from Google Trends, figure 6.1 shows whether the issue diffused to Google News or Twitter but figure 6.2 shows the rate of whole diffusion, including second node diffusion. From the analysis result, I need to have a look the trending topic diffusion process from Twitter. Unlike other diffusion process, if the trending topic starts from the Twitter, it is very hard to predict which direction the trending topic will be diffused by using only ‘seed service feature’.

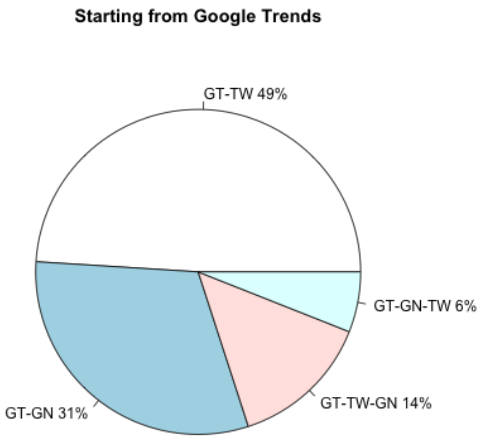

Fig. 6.1 Trending Topics Diffusion among three services - starting from Google Trends

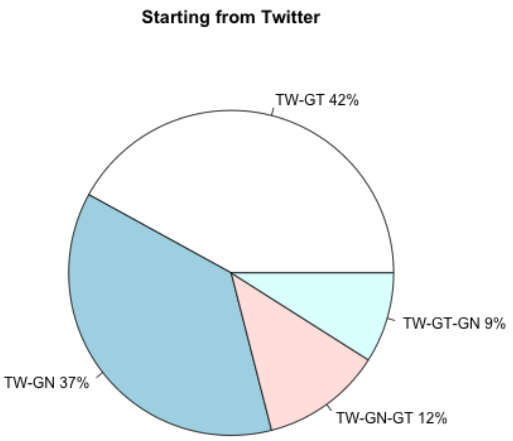

Fig. 6.2 Trending Topics Diffusion among three services - starting from Twitter

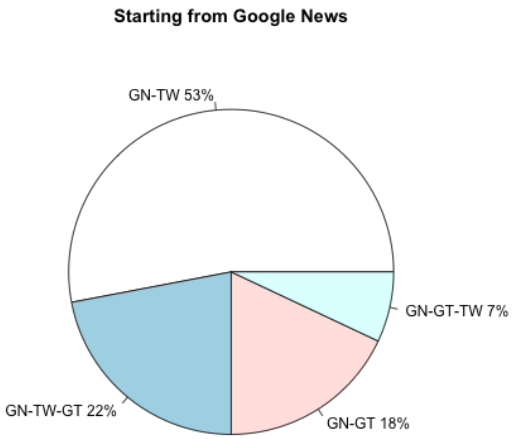

Fig. 6.3 Trending Topics Diffusion among three services - starting from Google News

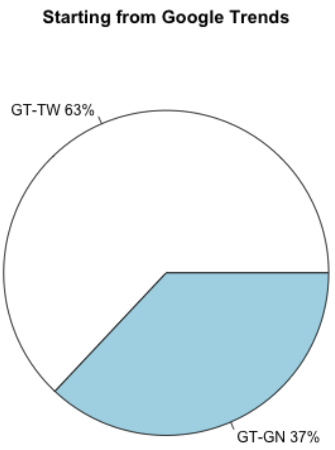

Fig. 6.4 Trending Topics Diffusion among three services - starting from Google Trends

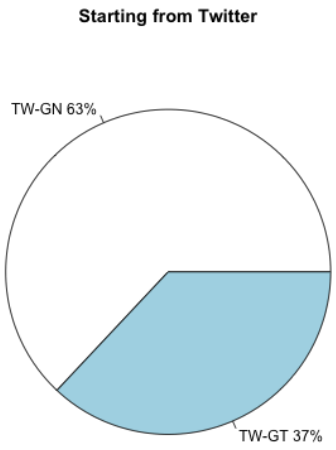

Fig. 6.5 Trending Topics Diffusion among three services - starting from Twitter

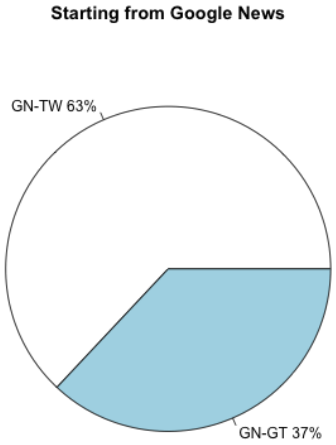

Fig. 6.6 Trending Topics Diffusion among three services - starting from Google News

6.4.2 Characterising of Time

Secondly, I explore whether the trending topic diffusion process is influenced to the time classes. According to the web information diffusion research (Morris 2012), time factor always affects the process of information diffusion on the web. It can be assumed that people

use web services on the morning have different interests from the users who use services on the evening or midnight.

Table 6.4 Trending Topics Diffusion based on the starting time

Google Trends	G-T	G-N	G-T-N	G-N-T	G only	common total	Total
- Night Time	25	19	6	3	290	53	343
- Business Time	37	21	15	3	449	76	525
- After Work Time	61	37	15	10	507	123	630
Total	123	77	36	16	1246	252	1498
Twitter	T-G	T-N	T-G-N	T-N-G	T only	common total	Total
- Night Time	118	97	23	24	3322	262	3584
- Business Time	194	183	39	67	7808	483	8291
- After Work Time	186	165	40	48	8297	439	8736
Total	498	445	102	139	19427	1184	20611
Google news	N-G	N-T	N-G-T	N-T-G	N only	common total	Total
- Night Time	23	94	11	27	1328	155	1483
- Business Time	55	148	24	70	1328	297	1625
After Work Time	31	85	9	40	1074	165	1239
Total	109	327	44	137	3730	617	4347

* G = Google Trends, T = Twitter, N = Google News

Time Classification

To support of the goal of characterizing timebased trending topic diffusion, I classify the time that trending topics are emerged into three different categories: Night Time, Business Time, and After Work Time. The time classes are divided based on the daily life of most people. As one day have 24 hours, I divide into three categories (8 hours):

- Night Time (1am – 9am): This time class can be called as ‘Early morning hours class”. People usually sleep in this period.
- Business Time (9am – 5pm): This time class can be called as ‘Office hours class’. People usually go to work or school from 9am to 5pm.
- After Work Time (5pm – 1am): This time can be called as “Evening hours class”. After people went to work or school, they usually take a rest in this time period. I can assume that the users using the services in the different time can have different interests in trending topic.

Characterisation of Time based Diffusion

To characterize the trending topic diffusion process based on the different time classes, I summaries the trending topic diffusion with each service and each time class in the table 6.4. Table 6.4 show each number of trending topic that started from certain service at certain time period. As can be seen the shaded part in the figure 6.4 , Twitter and Google News provides new trending topic keywords in the business time. It proves that Twitter and Google News users discusses or read issue or event in the Business Time. However, the nature of time of Google Trends is different from those two services. It normally started at the 'After Work Time'. Not surprisingly, all three services provides fewest new trending topic at 'Night Time (1am to 9am)'

6.4.3 Characterizing of Topic

Finally, I explore whether the trending topic diffusion process is affected by the topic of trending topic. Topic is the most popular features to be considered in information diffusion and information propagation prediction. The strategy can be used in trending topic diffusion process prediction.

Topic Identification and Classification

To support the hypotheses, I firstly need to identify and classify the topic for each trending topic from three services. For identifying the topic of trending topics, I decided the New York Times (NY times) classification service. This is because the trending topic from three services about the realtime events. Hence, traditional topic classification ontology cannot be applied. If the topic of trending topic are extracted from general document ontology made at anytime, semantically related topic will be extracted. Then, it may not obtain the exact meaning of the trending topic. The way to identify the topic of trending topic as follows: first of all, I search the trending topic keyword to NY times issue classification service. I set the published time as the day that trending topic firstly emerged. Then I can find the related article that published with that keyword at that day. As a result, NY times displays which category the trending topic is related. Based on the NY Times service, there are 9 different categories as below:

- Sports: For 'Sports', the trending topic should be about the sports games, athletes' names, and matching sports' name.
- Entertainment: For 'Entertainment', the trending topic should be related to the celebrities, art and cultures, travel, movies, books, and theater

- Politics: For ‘Politics’, the trending topic should be about the Politics names, and Parties.
- Business: For ‘Business’, the trending topic should be related to the economy, business, career and workspace field.
- World issue: For ‘World issue’, the trending topic should be related to the world issue (affected to the world)
- Technology: For ‘Technology’, the trending topic should be about technology, science, autos, and cars.
- Fashion: For ‘Fashion’, Style, dining, the trending topic should be related to home, lifestyleleisure, and serviceshopping.
- Obituaries: For ‘Obituaries’, the trending topic should be about crime, law, unrest, conflicts, war, disaster, and accidents.
- Health: To classify the trending topic to ‘Health’, the trending topic should be related to healthrelated news, flu, and infectious disease

Table 6.5 displays the topic distribution for each webservices. As you can see the table, there are three topics, which are always on top 3 categories, ‘Entertainment’, ‘Sports’, and ‘Politics’. The topic category ‘Entertainment’ is almost around 40% in both Google Trends and Twitter, however, Google News has similar distribution for all three categories. As a result, the main stream of trending topics for all three services is about Entertainment, Sports, and Politics

Table 6.5 Trending Topics Distribution based on the Web Services

Rank	Google News		Google Tends		Twitter Trending Topics	
	Topic	(%)	Topic	(%)	Topic	(%)
1	Entertainment	38%	Entertainment	42%	Sports	28%
2	Sports	25%	Entertainment	28%	Entertainment	26%
3	Politics	17%	Politics	10%	Politics	26%
4	Obituaries	5%	Fashion	6%	Health	7%
5	World issue	4%	World issue	5%	Business	4%
6	Fashion	4%	Obituaries	4%	Obituaries	4%
7	Health	3%	Health	2%	Technology	2%
8	Business	3%	Business	2%	Fashion	2%
9	Technology	2%	Technology	1%	World issue	2%

Characterisation of Topic based Diffusion

After I find out the topic distribution for each service, I examine whether the topic of trending topic is influenced the trending topic diffusion. From figure 6.7 to 6.9, I summaries the trending topic diffusion based on the topic of trending topics. The figure 6.7 to 6.9 shows the topic distribution of common trending topic keywords between two services, including G to N (Google Trends and Google News), G to T (Google Trends and Twitter), and N to T (Google News and Twitter). As you can see the figure 6.7 and 6.8, from the topic distribution in Google Trends to Google News and Google News to Twitter relationship, the top three topics (Entertainment, Sports, and Politics) are exactly same and the rate of distribution is also similar. Compare to services GN, GTTW relationship shows 44% is about sports. These differences can be helpful for predicting trending topic diffusion process.

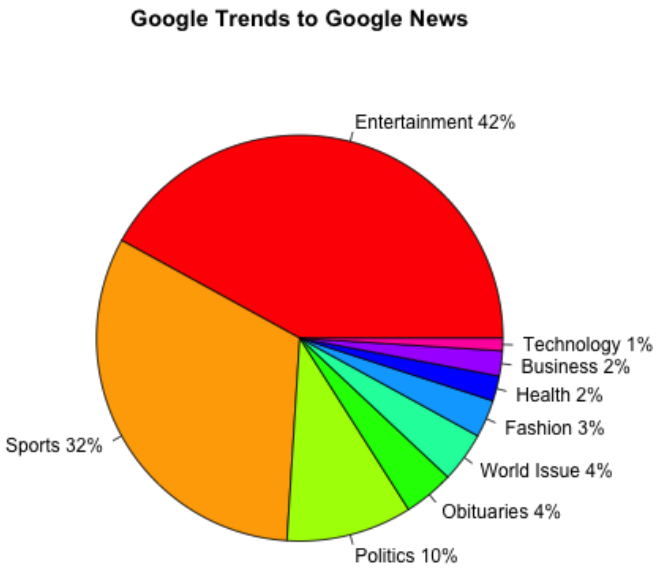

Fig. 6.7 Topic based Diffusion Pattern between Google Trends and Google News

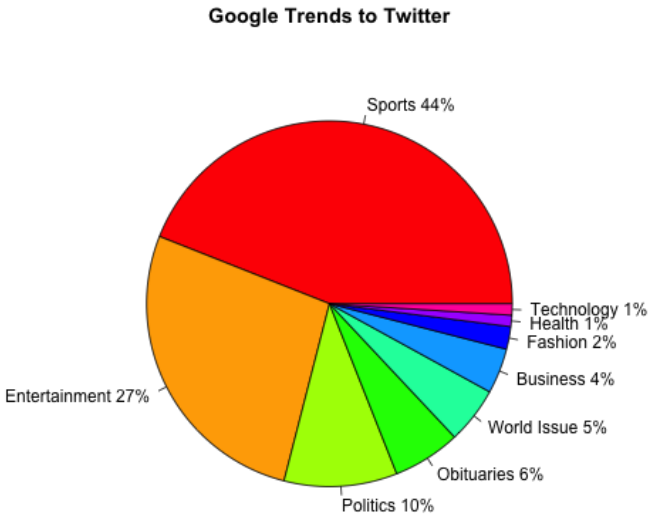

Fig. 6.8 Topic based Diffusion Pattern between Google Trends and Twitter

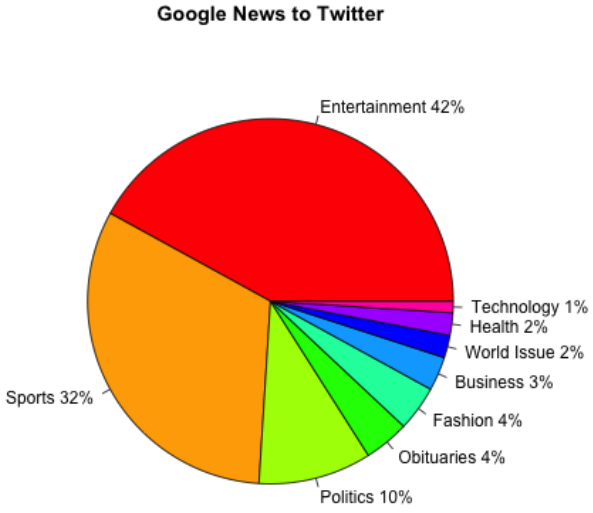

Fig. 6.9 Topic based Diffusion Pattern between Google News and Twitter

6.5 Evaluation Set-up

I describe the experiment result of prediction model. I conducted experiments to evaluate the prediction model for two different aspects: flow and periodicity of the trending topic. The experiment section is divided into two ways:

- Flow Prediction: Predict the next trending topic list provider that contains the keyword manifested in the seed provider
- Interval Prediction: Calculate the estimated the interval (delay time) between seed and following provider.

Both experiments follow the same algorithm to evaluate the prediction result.

Algorithm 2 Trending topics diffusion prediction

1: Extract a keyword k from the each service with extracting time t.
2: Classify the time t in to class C (Business time = BT, After work time = AW & Night time = NT).
3: For each keyword k, obtain a topic T (Sports = Sp, Politics = Po, Entertainment = En, Technology = Te, Fashion = Fa, Business = Bu, World issue = Wi, Health = He, Obituaries = Ob & None = No) of the keyword from real-time search in the online news site service.
4: Estimate the next trending topic list provider NP by applying machine learning approach
5: Calculate the interval (delay time) i between the seed trending topic list provider and the next provider by applying the same machine learning approach used in the process.

As described in the algorithm, this research applied several machine learning approaches to estimate next trending topic list provider and calculate the interval (delay time) between seed provider and next provider. There are eight machine learning approaches are used:

- Fuzzy: Fuzzy Unordered Rule Induction Algorithm
- SVM: Support Vector Machine
- IBk : K NN (K nearest neighbour)
- C4.5: C4.5 Decision Tree
- Kstar : K star
- Ridor : Ripple Down Rules
- MLP: FFNN (Feed Forward Neural Network)

- Logistics: logistic regression

This research employ above eight approaches with three proposed features and compare them using precision, recall and Fvalue of each feature

6.6 Evaluation Results

6.6.1 Flow of the trending topic

In the first experiment, the system predicts how the trending topic will be diffused among all (three) web services, including Google Trends, Twitter, and Google News. The system examines whether the keyword will be appeared on the other services, and if so which service will be the second or last. There are 6 possible predictions for each input keyword, as shown in the table 6.6. The sixth output (i.e. GT & TW) represents that the keyword will appear on the two providers top ten list in the same time.

Table 6.6 Output of Flow Prediction

	1	2	3	4	5	6
Google News (GN)	No Flow	GT	TW	GT ->TW	TW ->GT	GT & TW
Google Trends (GT)	No Flow	GN	TW	GN ->TW	TW ->GN	GN & TW
Twitter (TW)	No Flow	GN	GT	GN ->GT	GT ->GN	GN & GT

Table 6.7 shows results of the first experiment. I can see that SVM has the highest results when all three features are applied. When the seed provider feature is used, the results show better performance in general. It is apparent that the provider feature has better distinct characteristic than other features.

Table 6.7 Experiment result for flow of trending topics

Feature	Type	C4.5	Fuzzy	IBk	kStar	Logistics	MLP	RDR	SVM
Topic	Precision	0.512	0.578	0.535	0.51	0.523	0.521	0.503	0.475
	Recall	0.502	0.59	0.529	0.503	0.515	0.545	0.505	0.475
	F-value	0.513	0.563	0.53	0.517	0.529	0.528	0.495	0.475
Provider	Precision	0.534	0.623	0.563	0.621	0.603	0.633	0.635	0.641
	Recall	0.528	0.613	0.546	0.542	0.593	0.587	0.61	0.662
	F-value	0.53	0.618	0.552	0.589	0.61	0.608	0.612	0.649
Time	Precision	0.398	0.495	0.491	0.539	0.567	0.532	0.396	0.512
	Recall	0.631	0.587	0.484	0.595	0.611	0.563	0.623	0.599
	F-value	0.488	0.505	0.487	0.538	0.56	0.54	0.484	0.512
All	Precision	0.566	0.714	0.623	0.639	0.562	0.622	0.637	0.721
	Recall	0.752	0.737	0.613	0.736	0.734	0.686	0.747	0.757
	F-value	0.646	0.67	0.618	0.656	0.637	0.646	0.649	0.673

Table 6.8 Experiment result for interval between seed and following provider

Feature	Type	C4.5	IBk	kStar	Logistics	MLP	RDR	SVM	Fuzzy
Topic	Precision	0.328	0.311	0.386	0.308	0.337	0.327	0.328	0.452
	Recall	0.496	0.508	0.423	0.416	0.456	0.504	0.506	0.516
	F-value	0.376	0.356	0.321	0.338	0.376	0.372	0.344	0.413
Provider	Precision	0.308	0.313	0.307	0.312	0.328	0.304	0.414	0.402
	Recall	0.415	0.421	0.425	0.421	0.454	0.426	0.501	0.514
	F-value	0.332	0.405	0.335	0.376	0.345	0.326	0.33	0.422
Time	Precision	0.325	0.358	0.382	0.332	0.328	0.352	0.423	0.409
	Recall	0.456	0.396	0.443	0.422	0.496	0.487	0.436	0.476
	F-value	0.374	0.376	0.325	0.375	0.376	0.401	0.394	0.398
All	Precision	0.348	0.368	0.397	0.39	0.352	0.335	0.426	0.468
	Recall	0.457	0.396	0.455	0.456	0.487	0.487	0.428	0.554
	F-value	0.396	0.345	0.356	0.388	0.386	0.372	0.403	0.498

6.6.2 Interval between seed and following provider

In the second experiment, the system predicts how much time does trending topic requires being on the top 10 keywords list of other services. After the system estimates that the

keyword will appears on the other services, system calculates the time takes to be happened. As interval value range from the dataset varies, I normalized the time values. 15% of the time interval values are higher than 2 months and noisiness of data was relatively high to be used as it is. If the interval value is over 2 months, I consider it as “No Flow”. Within 2 month’s interval; I have grouped the interval values

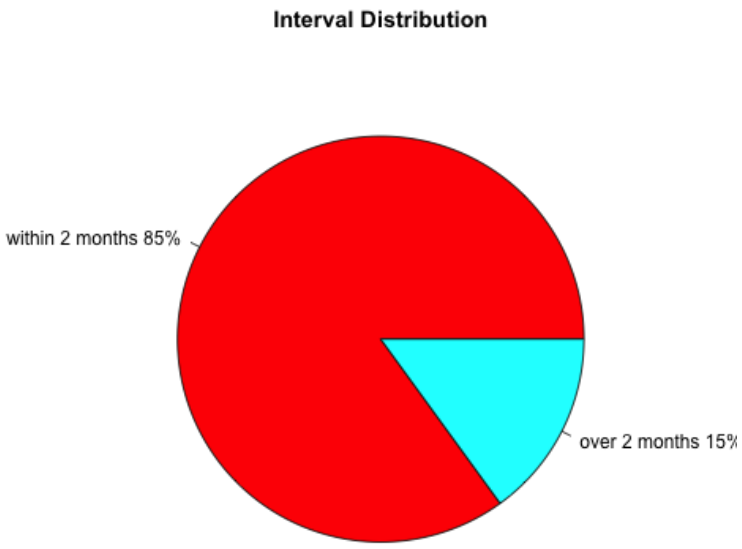

Fig. 6.10 Interval Distribution

Table 6.8 shows the experiment results of the interval prediction. The result represents that Fuzzy has the highest results when all three features are applied. Interestingly, when the seed provider feature is used, the results show that the performance is not much different from other features. This may mean that the distinctive characteristics of provider feature does not affect on interval prediction as shown in figure 6.10.

6.6.3 Additional Feature

Apart from the proposed features, there are several possible features can be considered to be added in the prediction model. Within a service, some trending topics may remain as popular keyword or become unpopular keyword. Thus, I examined the popularity of the trending topic in the each service by using rank of each keyword in the top 10 keywords list.

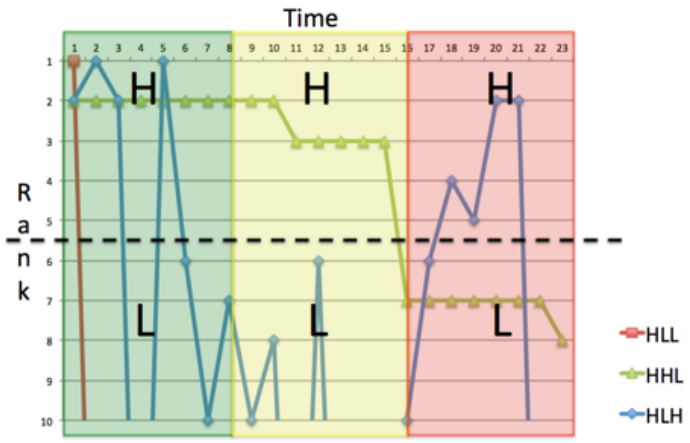

Fig. 6.11 24hour Trending Topics Rank Pattern

I examined 24 hours ranks (from first manifest) of each keyword to apply it on the prediction model. When the keyword is appeared on any of the target web services, I track the ranks of the trending topic on the top 10 lists for 24hours (from the first manifest). Then, 24 hours track is divided in to three phases (first, second and last 8 hours) and each phase is classified into H (High) or L (Low) class based on the highest rank within the each phase as shown in the figure 6.11. There are total eight types of 24hour rank patterns: HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL. The table 6.9 shows the distribution of the patterns among three services. This pattern will represent the trend of the trending topic for 24 hours after its first manifest.

Table 6.9 24-hours rank pattern distribution

	Google News (GN)	Google Trends (GT)	Twitter (TW)	Total
HHH	4.70%	31.80%	0.91%	7.21%
HHL	6.51%	19.82%	2.63%	6.43%
HLH	2.71%	0.46%	0.64%	1.41%
HLL	35.26%	21.20%	28.95%	30.30%
LHH	0.54%	0.92%	0.27%	0.44%
LHL	1.45%	0.46%	1.36%	1.32%
LLH	1.45%	0.92%	0.27%	0.68%
LLL	47.38%	24.42%	64.97%	52.22%
Total	100.00%	100.00%	100.00%	100.00%

In general, HLL and LLL are the most common patterns. The keywords manifested from Google Trends are likely to have more various patterns than others. The keywords from

Twitter are likely to have LLL pattern (64.97%). As shown in the table 6.9, each service has its own characteristic and this implies that the 24hour ranks pattern (RP) can be used as another feature of prediction model.

As can be seen in the table 6.10, I have added this feature in the prediction model and the result of the prediction accuracy rate has been increased as follows (All three proposed features are included in the below experiment):

Table 6.10 Experiment result for 24-hour ranks pattern prediction

	Flow Prediction_SVM	Flow Prediction_SVM with RP	Interval Prediction_Fuzzy	Interval Prediction_Fuzzy with RP
Precision	0.721	0.745	0.468	0.487
Recall	0.757	0.776	0.554	0.568
F-Value	0.673	0.693	0.498	0.519

24 hour ranks pattern feature obviously increases the performance of the prediction results, but it requires 24 hours delay in the prediction, as it needs to extract 24 hours ranks after the first manifests of keyword prior to the prediction. Thus, the keyword does not have 24 hours historic records cannot include this feature. When the prediction is made from the realtime data with dynamic dataset, this feature can be useful.

6.7 Implementation

The database for this research is designed for storing all detailed information of trending topics from different types (Google Trends, Google News, Twitter) of trending topics analytics services. It also contains the training data for the service diffusion prediction and accuracy results. The database for this project is designed as follows:

- Table: tb_twt_keyword
  - id: primary key, the identification number generated by auto increment function.
  - keyword: Twitter trending keyword
  - rank: rank of the Twitter trending keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - country: country of the Twitter trending keyword

  - local_time: local time at collection
  - date: collect time

- Table: tb_ggt_keyword
  - id: primary key, the identification number generated by auto increment function.
  - keyword: Google Trends keyword
  - rank: rank of the Google Trends keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - date: collect time

- Table: tb_ggn_keyword
  - id: primary key, the identification number generated by auto increment function.
  - keyword: Google News keyword
  - rank: rank of the Google News keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - date: collect time

- Table: tb_twt_keyword_distribution
  - id: primary key, the identification number generated by auto increment function.
  - twt_keyword: Twitter trending keyword
  - active_period_id: identification number of the current keyword active period
  - first_appearance_group: identification number of the current keyword active period
  - last_appearance_group: identification number of the current keyword active period
  - life_time: life time of the current keyword active period

- Table: tb_ggt_keyword_distribution
  - id: primary key, the identification number generated by auto increment function.
  - twt_keyword: Google Trends keyword
  - active_period_id: identification number of the current keyword active period

  - first_appearance_group: identification number of the current keyword active period
  - last_appearance_group: identification number of the current keyword active period
  - life_time: life time of the current keyword active period

- Table: tb_ggn_keyword_distribution
  - id: primary key, the identification number generated by auto increment function.
  - twt_keyword: Google News keyword
  - active_period_id: identification number of the current keyword active period
  - first_appearance_group: identification number of the current keyword active period
  - last_appearance_group: identification number of the current keyword active period
  - life_time: life time of the current keyword active period

- Table: tb_training_data
  - keyword: keyword
  - topic: topic of the keyword
  - time_class: time classification of the keyword
  - rank_pattern: rank pattern of the keyword
  - diffusion_route: service diffusion route of keyword

- Table: tb_prediction_results
  - tb_training_data_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweets table
  - accuracy: diffusion prediction accuracy of the keyword

The summary of database design can be found in the Appendix A.

6.8 Conlcusion

As mentioned in this research, I investigated trending topic lists in the web-services, such as Google Trends, Google News and Twitter Trending Topics. I consider that three features, including seed provider, time classification and topic of the keyword, have its own characteristics and can increase the prediction performance for trending topic diffusion process. I have analyzed each feature and evaluate the usefulness of the feature in the prediction model of trending topic diffusion by applying different machine learning approach. I have also found that the rank patterns in the trending topic list can be used as additional feature. It is hoped that this research provides some sights into future research of trending topic diffusion modeling on the web.

Chapter 7

Trending Topics Diffusion Prediction Among Countries

Chapter 8 introduces trending topic diffusion prediction model among online communities in different countries.

7.1 Introduction

In decades ago, people received information passively from offline media, such as newspaper, television and face-to-face interaction. Nowadays, people share and get information from online services, which enable individuals or organizations to connect. People always talk about lots of topics as figure 7.1 shows. Some of these topics become popular topics as the blue colour of topics goes to trending topics list that can be provided by some online services. These online services, such as social media, search engines, online news, and other like Wikipedia, collect their users' social data to figure out what people are currently most discussed about.

Under their monitoring, the data social data collected from these web services has strong persuasion to represent the people's interestingness online. For example, people pay attention on the online platforms, called social media that enable users post or read messages in short text, which almost have millions of registered users. Twitter enables people to share their interests, activities, backgrounds and real-life connections. It monitors and analyses these social data to extract the most 10 popular topics with ranking for real time in trending topics list. Since Twitter has amount of users whose behavior has representativeness, and trending topics represents the topics users are most interested in. Trending topics provided by Twitter reflect the popular topics people are currently interested in. Kwak et al. (2010) indicated that

the over 85% trending topics from Twitter are headline news or persistent news in nature, which means majority of trending topics from Twitter associated with real world. Trending topics can represent the real world events, so it might be valuable for researchers to uncover the people's interests around world.

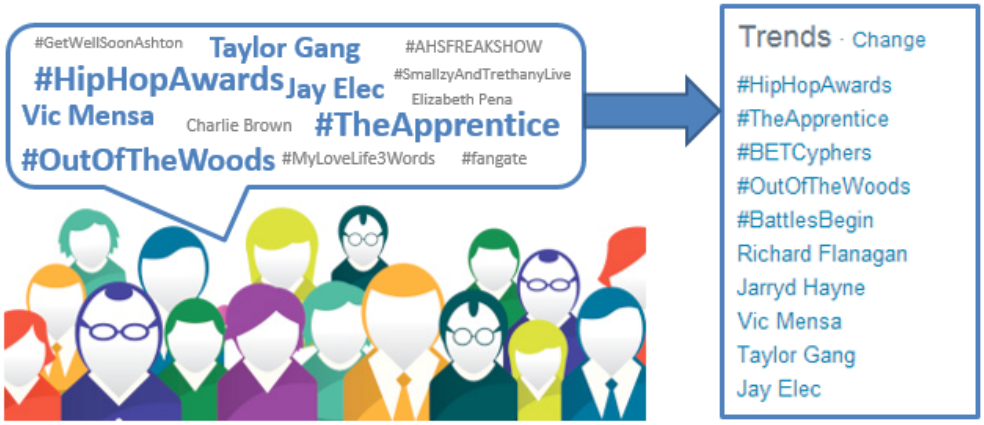

Fig. 7.1 Trending Topics Circumstance

Twitter trending topics are displayed on the middle-right side of the Twitter interface, and those are from several small cities to worldwide. Hence, these topics represent from local to worldwide events. The interesting phenomenon we found from trending topics is that different trending topics appeared in various kinds of cities and countries and some of them diffused. This phenomena can be seen that trending topics spread into countries. The spread of trending topics can be regarded as the diffusion of trending topics from a place to others. As we can see in figure 7.2, this map shows the diffusion of IOS 8 among different countries since the release of IOS 8. Trending topic of IOS 8 is started in USA, and then goes to Canada and UK, and then goes to Malaysia, Philippines and Australia, at last it diffused into New Zealand and Singapore. Once tracked trending topic of IOS 8, it is obvious that IOS 8 diffused among these eight countries. In fact, by analysing actual trending topics data from Twitter, a great amount of trending topics are diffused as table 7.1 shown. Table 7.1 shows the analysis result of the three-month period tweets containing real-time top 10 trending topics that provided by Twitter. The number of appeared trending topics in table 7.1 represent that trending topics appeared in each country. The number of diffused trending topics in table 7.1 represents that the trending topics appeared in multiple countries. The percentages of diffusion for each country are calculated as follows: the number of diffused trending topics divides by the number of appeared trending topics in each country. As table 7.1, over 90% of topics in eight English-speaking countries are diffused among different countries. The result shows a phenomenon that most trending topics appeared in multiple

countries, which demonstrates the diffusion of trending topics exists and it is important to research.

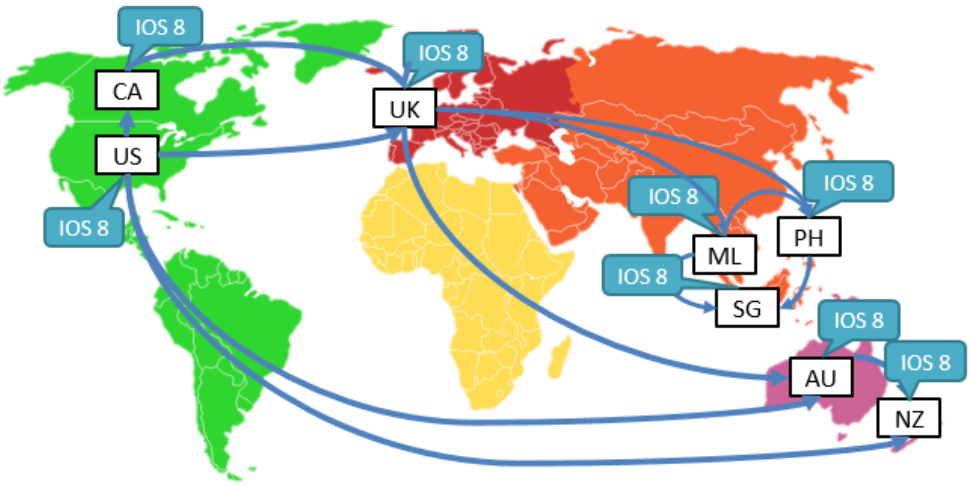

Fig. 7.2 Diffusion of iOS 8 across 8 different english speaking countries

The diffusion of trending topics represents people's interests among countries, which can used to detect activities in real world. The diffusion of trending topics can be captured into three major properties: speed, scale and range. The speed of diffusion of trending topics represents how long a diffusion of a trending topic takes. The scale of diffusion of trending topics means how many countries a trending topic diffuses. And the range of diffusion of trending topics is to investigate how far the diffusion chain of a trending topic can continue on in depth. This research aims to model a prediction diffusion to predict the diffusion trends of trending topics including scale and range of diffusion of trending topics. If we can predict the diffusion trends of trending topics among different countries, it can be used in economic and financial purposes. For example, once a trending topic are tracked, the diffusion of trending topic can be predicted, how many countries the trending topics will diffuse, which can be used in marketing research of products. As the previous diffusion example of IOS 8, which help Apple Inc. use this track to do some marketing research about popularity of IOS 8, such as predicting the sales in different countries or designed different functions for people in different countries. Another example is that if trending topics about security issues always diffused into USA, Malaysia or Philippines, then the international companies should pay more attention on the security department in these countries than others.

In order to model a prediction diffusion, it is necessary to classify trending topics by different characters. Trending topics can be classified by the prediction diffusion with country factor, context factor and ranking factor. The country factor represents the characters of the countries that starts a trending topic, which has two extracted features including diffusing

level feature and speaking language feature. Through analysing the actual trending topics data, it is found that these eight English-speaking countries play different roles when trending topics diffusing. Some countries are mostly started trending topics, and some countries are mostly discussing about trending topics that are started from others. Moreover, the starting countries has another character about speaking language. Although these eight countries speak English, the language can be divided in to western English and eastern English. Such as USA and UK speak western English while Philippines and Singapore speak eastern English. Countries speaking different languages pay attention on different trending topics. The second one is the context factor represents the character of context of trending topics. In spite of various context of trending topics, these trending topics can be classified into several categories by context patterns. For example, trending topics as Be + noun differ with trending topics about holidays as Christmas. In order to classify context of trending topics, the rule tree is made with context patterns. The third factor is the ranking factor utilizing the ranks provided by Twitter, which represents the popularity of trending topics. Due to each rank provided by Twitter represents the popularity of a trending topic in one countries at one time, these ranks for each trending topic have to filter and keep the applicable ranks. Due to there may be several starting countries for a trending topic, the different ranks of a trending topic should be integrated into one rank. The average ranking is to average the ranks of a trending topic appeared at first time in multiple countries and categorized it into relevant level. In contrast, starting ranking feature is to consider the diffusing level of starting countries for a trending topic into it.

This research focuses on building a prediction diffusion and predicting the scale and range of diffusion of trending topics. To model prediction diffusion, the actual trending topics data that collected from Twitter from 8th August 2013 to 07th November 2013 in eight English-speaking countries (the United States, the United Kingdom, Canada, Australia, New Zealand, Philippines, Malaysia, and Singapore) is analysed, which includes 3975 unique trending topics keywords in 376077 tweets that containing these trending topics. The contribution of this chapter is the initial paper of modelling and predicting the diffusion trends of trending topics that to track how many countries that a trending topic diffuses and how far the diffusion chain of a trending topic can continue on in depth. And I analysed three month actual trending topics data provided by Twitter to model a prediction model with country factor, context factor and ranking factor.

7.2 Related Work

In order to build an information diffusion model for the online trending topics, the background should be distinct into four parts which contains how the trending topics can be identified, the history of information diffusion, these information diffusion models and how these researchers do the researches, and how researchers applied social media data into prediction.

7.2.1 Trending Topics Diffusion

The context is easy to understand from their literal meanings. The identification of trending topics describes how the previous researchers extracted data for data collection. The trending topics describes extraction of trending topics and utilization of prediction of trending topics. The information diffusion modelling describes the approaches to model information diffusion and existed information diffusion models, and how researchers use the information diffusion for their researches.

With the development of network, websites and online applications provide information that people may interested in. However, the data is too big to find out the real-time information or previous information, which requires researchers to extract the news from diverse webpages or datasets. As early as 1992, Andersen et al. proposed JASPER which applying template-driven method to extract news for solving significant business problems.

The original intention of JASPER is to help the Reuters to analyse the financial news and reports which are provided by publicly-traded companies. Once the earnings and dividend reports are generated by JASPER, the reporters only need check the necessary information, which helps the decision-makers make better decisions fast and accurately. Although Andersen et al. (1992) extracted information from text and table, they mentioned JASPER has a frame representation to check whether the new PR Newswire releases match the patterns and decides to assign a value to the slot. JASPER generates a new story from these information which is available for reporter to edit after the extracting and storing all available information.

Andersen et al. (1992) try to evaluate the accuracy of extracted information for earnings and dividend releases provided by JASPER. They considered the measures of accuracy as completeness and correctness by testing 100 earnings and 50 dividend releases. Similarly, in order to extract relevant content, Laber et al. (2009) proposed NCE (News Content Extractor) to work. They indicated that their method works based on DOM tree representation of new web pages, which also applied in research of Reis et al. (2004) that extract information automatically from websites. Laber et al. (2009) assumed two hypotheses, which there is high measure of a node associated with the webpage and a positive real number, and which

comments display after body of a news webpage. According to their observation, the measure of a news webpage achieves almost 90% by testing 324 news documents.

However, in research of Reis et al. (2004), they rather study a specific type of tree called labelled ordered rooted tree. They presented a new algorithm for determining a new type of mapping called RTDM (Restricted Top-Down Mapping), which extracts information by page clustering, extraction pattern generation, data matching, and data labelling. They compared the extracted news by original HTML pages and by their approach from 35 sites. And the average 87.71% correctly results demonstrate the RTDM algorithm has highly effective for extracting new automatically.

Xia, Yu & Zhang (2009) also applied tree alignment algorithm for their research, proposed an automatic wrapper generation method. A heuristic method is employed for determining the most probable content block and the alignment algorithm detects repeating patterns on the union tree. Therefore, they compared their approach to RTDM which applied in research of Reis et al. (2004) by testing 9000 web pages including Blog, news, forums. The results show that the performance of proposed approach in blogs website is better than in news and forums websites. Although the results of their new tree alignment display out performances in Blog, news and forums website, the extracted information is complex and comprehensive. Ma & Wan (2010) provided an approach to classify only news comments from readers.

Their approach aims to extract explicit and implicit opinion targets from news comments by based on Centring Theory. Ma & Wan (2010) extracted 'focused concepts and rank their importance by computing the semantic relatedness with sentences via Wikipedia'. The experiment demonstrates that the approach effective. However, their results are not obvious high accuracy. The information extraction not only are news websites, forums and Blog, but also can be used in social media side such as Twitter. Medvet & Bartoli (2012) proposed an approach to detect popular topics, summarize these topics by the representation of their precise meanings, and evaluate sentiment polarity of each topic. Their approach employed with a given topic, which means they should collect the recent tweets related to that topic. After data collection, they identified the high quality of tweets and classified these tweets into three sentiment categories (positive, negative, or neutral). And then these representation tweets are summarized for each sentiment categories by Medvet & Bartoli (2012). They tested their approach to explain the precise meaningful qualitative evaluation of popular topics.

7.2.2 Information Diffusion Modeling

The early diffusion models are developed by Bass Diffusion model that consists of a simple differential equation to describe the process of how new products get adopted in a population

(Bass diffusion model 2014). This model shows the relationship between the current adopters and potential adopters with a new product interact. The adopters can be divided into innovators or imitators, and the speed and time of adoption depend on the degree of innovation and degree of imitation. The Bass model has been widely used in forecasting area, especially in new products' sales forecasting and technology forecasting.

However, Roger's published Diffusion of Innovations to against a typographical error of Bass paper, which is a highly influential work that described the different stages of product adoption. Diffusion of innovations is attempt to figure out how, why, and at what rate new ideas and technology spread through cultures. There are four main elements that influence the spread of a new idea: the innovation, communication channels, time, and a social system for this theory (Diffusion of innovations 2014). With the growth of rate of adoption, there is critical mass when an innovation researches saturation. According to (Rogers 1962), the adopters can be categories as five categories which containing innovators, early adopters, early majority, late majority, and laggard. As we can see, table 7.1 presents the definition of each category of adopters, which is proposed by Roger (1962).

Table 7.1 Information Diffusion Level - Roger(1962)

Adopter type	Definition
Innovators	Innovators are the first individuals to adopt an innovation
Early adopters	Early adopters are the second fastest category of individuals who adopt an innovation
Early Majority	Early Majority are individuals adopt an innovation after a varying degree of time
Late Majority	Late Majority are individuals adopt an innovation after the average member of the society
Laggards	Laggards are the last to adopt an innovation

According to Rogers, diffusion is a process by which an innovation is communicated through certain channels over time among the members of a social system. However, Foster et. al (2009) thought that diffusion is a process by which information, viruses, ideas and new behavior spread over social networks. The existed work is based on (Granovetter 1978)'s initial treatment of the phenomenon of collective behavior in 1978. He introduces a threshold model and uses it to examine the occurrence of riots and their perceived domino-effect growth pattern. The result was an early threshold model for collective behavior. Nowadays, diffusion models of social networks have been studied in a variety of fields ranging from epidemiology, to marketing, to technology transfers, to computer virus transmission, and to power systems.

Kempe et. al (2003) proposed the two basic diffusion models including the linear threshold model and the independent cascade model. The linear threshold model is for

which a node becomes active if a predetermined fraction, called a threshold, of the node's neighbours are active. In contrast, the independent cascade model highlights whenever a node becomes active, it gets a one-time chance to activate each of its neighbouring nodes with some probability. The rich literature for both models of existed work, especially the independent cascade model attracted people's attention in present day (Rogers 1962).

In terms of the phenomena of market to consume a new product, Bass found that although the new adopters reach the saturation point eventually, the innovators of new product are declining among the time of consumption of new product. At the meanwhile, number of imitators are accelerated from zero to peak value after a while of adopting the new product. And then the imitators of new product are decreasing after the peak point among the time of adopting new product. This interested phenomena almost the universal in all of adopting new product or new technology in market. In hence, Bass proposed an equation with the coefficient of innovation and coefficient of imitation (respectively represent external influence and internal influence) for the adoption of new production or new technology.

According to basic researches, the bass diffusion model is the original model to present the diffusion of information, which is used in long-term sales forecasting pattern of new product or new technology at beginning. Lilien, Rangaswamy& Bruyn (2007) thought that in general, this model is used under one of these two conditions. Either, the new product or new technology has been introduced into market and its sales has been observed for a short period. Or, the new product or new technology hasn't to be introduced into market, but there is already some existing analogue product or technology in market. And the sales pattern of these existing analogue product or technology already known. According to Roger (1962), while this model use for new product's sale forecasting, the prediction of this model should be focused on the number of customer will adopt this new product and when they will adopt it. Roger (1962) categorized these adopters by time into innovators, early adopters, early majority, late majority, and laggard, which will help the companies to leverage the resources.

Based on bass diffusion model, Lilien, Rangaswamy& Bruyn (2007) extend several key assumptions to provide the framework for the modelling the time path of new product or new technology adoption. They utilized past data to present sales pattern for analogue product to forecast the diffusion of a new product and make a decision for investment. Bass model provides the equation for the sales forecasting of a durable product traditionally in market. However, the diffusion models have become more and more complex while this theory is utilize more and more frequent. Radas (2005) put forward that to incorporate the influence of marketing mix variables as the external influences is the notable challenge for diffusion modelling. In hence, she summarized different models and contrast their advantages and

disadvantages. She identified two approaches to incorporate the marketing mix variables which containing the pre-specified way and parameters changes by time.

She utilized the real monthly sales data of existing durable product to verify that the original bass model need to incorporate external influences to forecast. Several researchers generated extension of bass model due to limitation of bass model which not incorporate marketing mix variables. Radas (2005) summarized the common marketing mix variables are price and advertising after reviewing other researchers' work. However, these marketing mix variables are constant, while the market change all the time. And the constant parameters are not enough to show the diffusion. In hence, some researchers proposed that allowing the parameter to vary among the time. Despite of the assumptions of the diffusion models, the goal of these models is to provide the flexible and easy way for better decision making of managers.

Table 7.2 Diffusion Prediction Approaches Summary

Purpose	Data	Approaches	Applied research
Forecasting the diffusion of focal product	Sales forecasting of analogous products	nonlinear regression, maximum likelihood estimation, Hierarchical Bayes estimation	Lilien, Rangaswamy& Bruyn 2007
Providing a framework for systematizing diffusion models focusing on factor marketing mix variables	Monthly sales data of an existing durable product	Constant parameters, parameters change by time	Radas 2005

With the development of network, people pay more attention on the e-life. The new lifestyle also attracted researchers to investigate the information diffusion process. At the beginning of 20th century, blog, as the new discussion and informational site for posting new articles, pictures and videos to express emotion and share information, was the most popular online life for people, especially young people. Although blogs can connect to each other through by blogrolls, comments, linkbacks and backlinks. In general, the popular dynamics are citation and affiliation. However, citations, which represent the value of the blog to cite while people reading it, are more indicative than affiliated. Therefore, researchers attempted to propose the information diffusion models to calculate the possibility of a post will cite or link to another one, or how a post diffuse from a blog to another blog. Two basic information diffusion models proposed to transform to different information diffusion models in variety of ways different researcher's' own assumptions. The existed works are covering most models

of information diffusion, in terms of the different situations the models and approaches are diverse. Kwon et al. (2009) assumed that relationships the between blogs present explicit relationships, while relationships of a blog between posts present trackback or scraped posts. Lim et al (2011) assume that all of the blogger actions are explicit, and they try to use the algorithms to assign the a diffusion probability to an edge for a pair of bloggers by defining the score of blogger as weight of blogger actions multiply each counts of blogger actions. Kwon et al. (2009) put forward the concept of the super node to diffuse by analysing the 100 million posts.

The proposed basic models are presented in term of the previous work (Kwon et al. 2009; Lim et al. 2011; McGlohon et al. 2007; Leskovec et al. 2007) and the researchers modify the model to suit for their researches. All of researches studied by Kwon et al. (2009), Kwon et al. (2009), Lim et al. (2011) and McGlohon et al. (2007) proposed the algorithms for calculating the diffusing probability of the post through nodes in blog network. Even Lim et al. (2011) presented the basic equation for calculating the probability of post A diffusing to post B. In order to distinguish the two different intentions of bloggers, they defined score of a post as the degree of intentions of diffusing the post of bloggers. Similarly, Kwon et al. (2009) adopted the algorithms to calculate the diffusion probability of the posts in the blog network. However, the algorithms of Kwon et al. (2009) are based on the new elements containing a super node, broadcast edges and register edges, which differ from the equation of Lim et al. (2011). Besides, Leskovec et al. (2007) and Goetz et al. (2009) found temporal patterns and topological patterns of blog, and presented respectively a generation model and zero-crossing model. They presented that the probability of the cascades of nodes follows a Zipf distribution by observing the degree distributions of the cascades. Each research has its own proper method and model with its specific conditions.

In blog world, the blog network between the bloggers of cascading of the posts and the cascading of post between nodes. McGlohon et al. (2007) developed the model of cascading behavior based on the threshold model and the cascades of posts from a blog to another and through nodes in the network. However, Lim et al. (2011) applied independent cascade model after clearing the blogosphere network to combine the probability of the post from A diffuse to B and the score of diffusing intention of the a post in blog A. This conceptual model presented by Leskovec et al. (2007) can produce cascade graphs which match the characteristics of realistic cascades. The model was proposed to figure out the cascades of nodes getting infected. At beginning, two states of blog are infected nodes and susceptible nodes. Based on the outcomes from the temporal patterns and topological patterns of blog to verify whether the generation model match the degree of distribution.

7.3 Methodology

7.3.1 Feature Selection

The aim of this research is to predict the diffusion trends that contain scale and range of trending topics. In order to predict these diffusion trends, the prediction model should be built. The following will present the proposed factors for the prediction diffusion and explain these factors in specific way.

Trending topics are most popular online topics people are interested in. Online social services provide platforms for global users to post messages. As an example, there is video in Twitter that a young people in Australia who bought the iPhone 6 immediately after the release of iPhone 6 show his new phone and excited emotion and then he dropped this new phone without a glance. And some of others in other countries like in Asia gloat the guy who dropped iPhone 6, some of them sympathize him. Although dropping iPhone 6 happened in Australia and the video posted in Australia, this topic can be diffused to other countries.

In order to research the diffusion trends of trending topics, the prediction diffusion needs to build. In order to conduct the prediction diffusion of trending topics, the actual three month trending topics data was analysed from 8th August 2013 to 07th November 2013 in eight English-speaking countries (the United States, the United Kingdom, Canada, Australia, New Zealand, Philippines, Malaysia, and Singapore), which includes 3975 unique trending topics keywords in 376077 tweets that containing these trending topics. For Twitter, trending topics can be extracted by monitoring and analysing users' social data. Trending topics data collected from Twitter every 15 minutes, which is composed of a trending topic that are extracted from the top ten real-time trending topics list provided by Twitter, the country that appeared this trending topic, the exactly time (shift the local time of each country into Australia time) that this trending topic appeared and the rank of this trending topic when it appeared.

Twitter published different trending topics in small city to worldwide. Therefore, different trending topics are appeared in different countries. And trending topics diffused among different countries. For example, the analysis result presents the percentage of trending topics appeared in USA that also appeared in other countries. It is obvious that the percentage of trending topics appeared only in USA are much less than the percentage of trending topics appeared in USA and other countries. As we can see in figure 7.3, 26% of trending topics appeared both in USA and UK, 33% of trending topics appeared both in USA and CA, and 11%, 10%, 7%, 5% and 4% present respectively the percentages of trending topics appeared both in USA and AU, PH, ML, SG and NZ. There are only 4% of trending topics appeared only in USA, which means 96% of trending topics that appeared in USA diffused. Similarly,

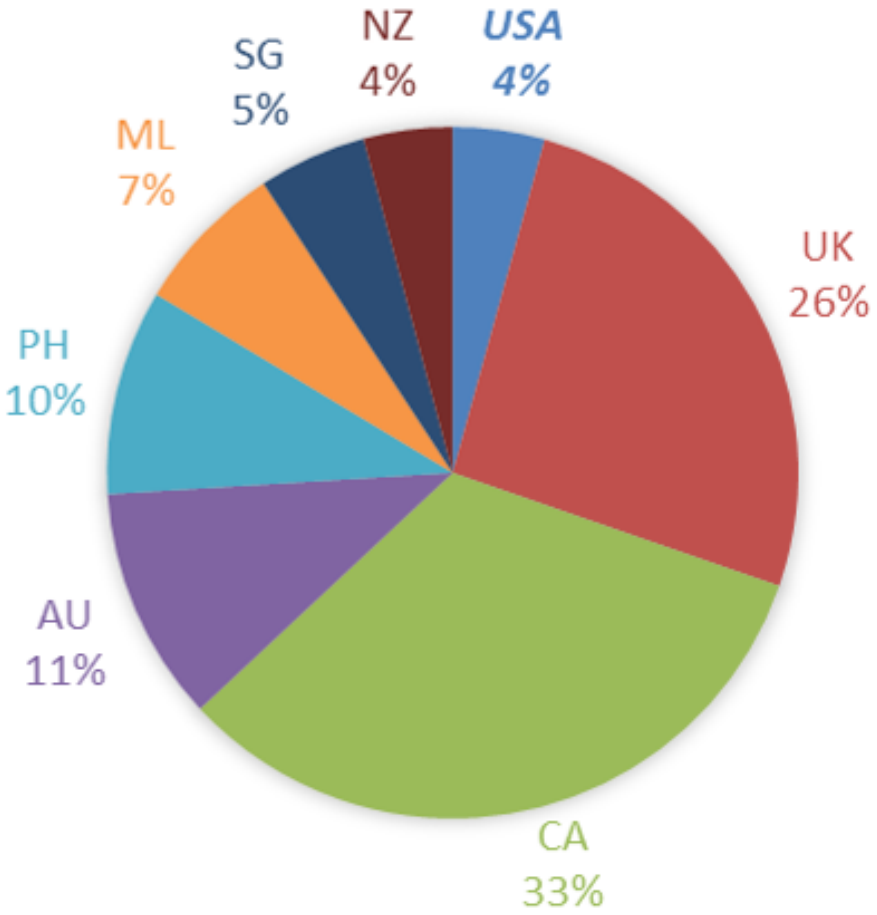

Fig. 7.3 Percentage of trending topics appeared in USA and diffused to other countries

the percentages of trending topics that appeared in single country are less than 5%, almost around 1% to 3%. The percentages of trending topics that appeared in multiple countries are much more than appeared in single country. These figures demonstrate that most trending topics diffused among multiple countries, which means the diffusion of trending topics is important to research.

Country Feature Diffusion Level

Just as its name implies, country factor represents the characters of the countries that start trending topics. The eight countries have different characters due to the different culture and background, which can be concluded into the different features. Due to these differences, people in different countries are interested in different trending topics. For example, trending topics that started in USA always are news about athletes and matches, in contrast, trending topics that started in ML are more about celebrating a memorial day like someone's birthday.

That's means trending topics may be different while they started from different countries. By analysing actual trending topics data, the diffusing level feature for starting countries and speaking language feature for the country that started trending topics are extracted by considering the characters of countries. Since trending topics could appear in different countries at same time. The country appeared trending topics at the earliest time is regarded as the starting country. For the collected trending topics data, the starting country of a trending topic can be several. Namely, each trending topic has a starting country or several starting countries, the starting countries play different roles in the diffusion of trending topics. There are two features for country factor, which are diffusing level feature and language feature that are described in specific way in follows.

In the diffusion of trending topics, some countries may start most of trending topics actively, some of them may receive trending topics from other countries actively, some of them may not care about trending topics appeared in other countries. Since different countries may play different roles in the diffusion of trending topics, diffusing level feature presents the role of country that started a trending topic in the diffusion of trending topics. For example, UK always starts trending topics and diffused them to US and CA, oppositely, North Korea never care about trending topics that discussed in other countries.

A trending topic appeared in multiple countries represents this trending topic diffused. So the percentage of diffused trending topics is same as the percentage of trending topics appeared in multiple countries. The three month actual Twitter trending topics data from eight English-speaking countries is analyzed. The analysis result shows that over 90% of trending topics appeared in multiple countries as figure 7.4 shows, which represents that most of trending topics in eight counties diffused. Although eight countries are caring about topics that are appeared in other countries, it has chance that there may be a country don't care about trending topics that started and diffused in other countries, such as North Korea who always ignores topics discussed in other countries. According to the percentage of diffused trending topics, countries can be classified as ignorant and diffuser. Ignorant refers to country always ignore trending topics that appeared in other countries. Diffuser refers to country started or diffused trending topics that appeared in other countries. Since there are two categories, the criteria to distinct ignorant and diffuser is whether over 50% of diffused trending topics in designed country.

In order to name three levels appropriately, Roger's categories of adopters for an innovation are considered. At early as 1962, Roger proposed that the adopters of an innovation can be categorized into five types which composed of innovator, early adopters, early majority, late majority and laggards. He (1962) defined the innovators as the first class to adopt the innovation; the early adopters as the second fast individuals to adopt the innovation; the

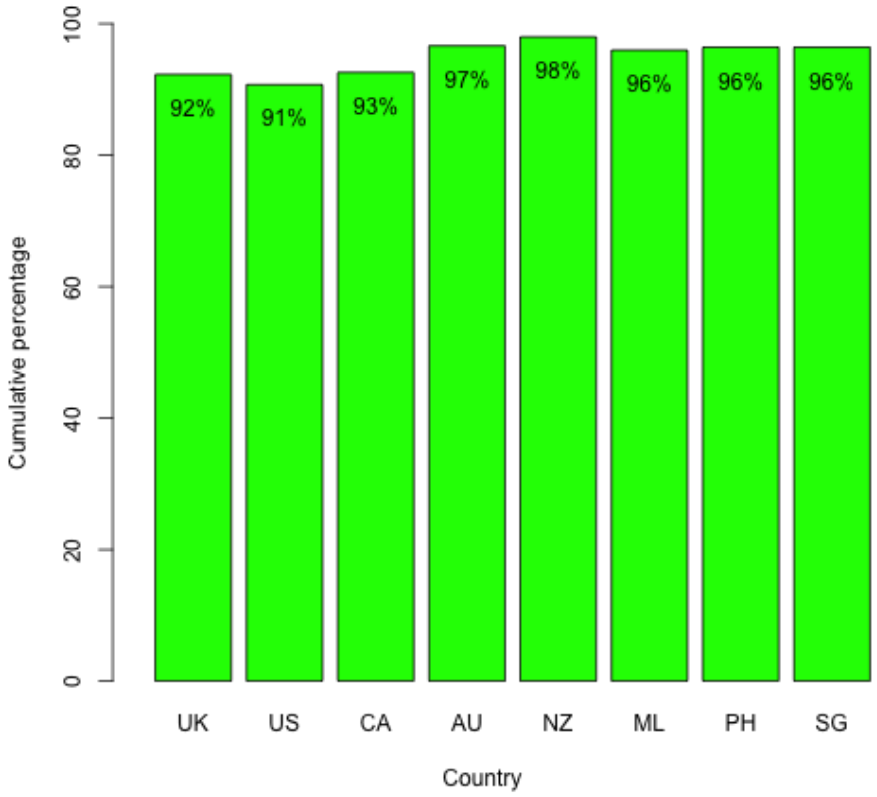

Fig. 7.4 Percentage of trending topics appears in multiple countries

early majority and late majority as the major individuals to adopt innovation, which are distinguished by their social status; and laggards as the last category of individuals to adopt the innovation. The initiation of a trending topic in this research is similar to adopting an innovation from Roger (1962). Therefore, the transformation of Roger's adopter categories for the diffusing level of starting countries applied in to this research is identified into three categories containing innovator, adopter and ignorant as table 10 shows. In order to adapt for this research, the definition of innovator, adopter and ignorant should be re- identified. The innovator for diffusing level refers to the country starts the trending topics and diffuses these trending topics to other countries mostly. And the adopter for diffusing level refers to the country receives the trending topics from innovators and diffuses them to other countries mostly. The ignorant for diffusing level refers to the country always ignore trending topics that started and diffused in other countries.

Based on above analysis, the results for categorizing these eight countries into different diffusing level present as table 11. Since these eight countries have over 90% of trending

topics diffused, which means all of them are diffusers. USA and UK are over the threshold for classifying and accord with the definition of innovator, which are identified as innovator. At the meanwhile, the percentages of that trending topics for CA, AU, NZ, PH, MA and SG have obvious weakness in diffusion initiation, which are identified as adopters. After applying this feature to trending topics data, the percentage of diffusing level of trending topics shows as table 12. 70% of trending topics are classified into innovator of diffusing level, only 30% of trending topics are classified into adopter of diffusing level. The high percentage results of innovator represents that innovators such as USA and UK affect the diffusion of trending topics a lot.

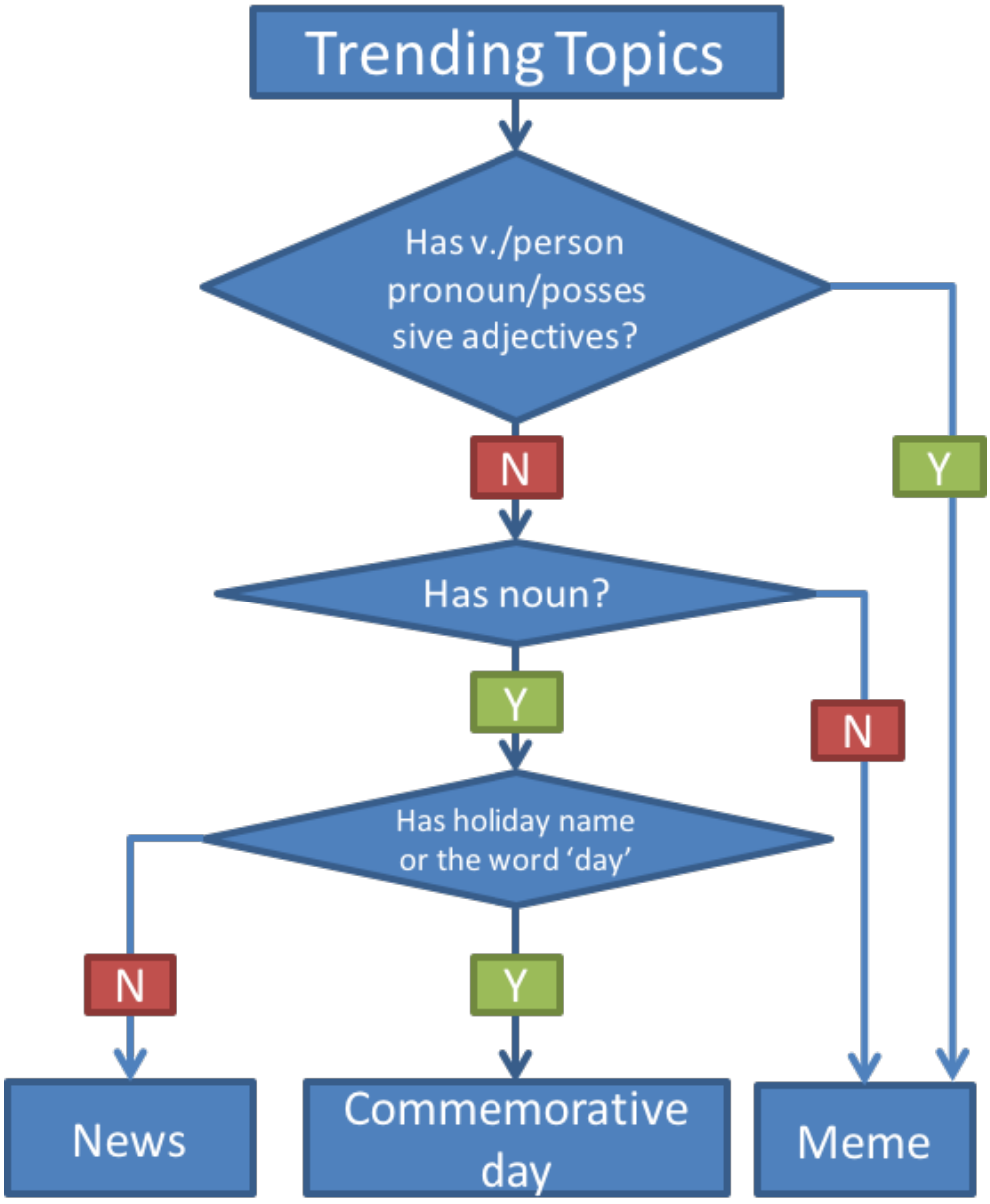

Fig. 7.5 Context feature Pattern Classification using Rule

Table 7.3 Percentage of trending topics based on the context pattern

Context categories	Percentage
Commemoratives	4%
Meme	9%
News	87%

Context Feature - Context Pattern

For a trending topic, there are several attributes including trending topic, appeared country, appeared time and rank. Appeared country for a trending topic is used for country factor. Similarly, trending topics itself can be used to classify. Context factor represents the character of context of trending topics. In this chapter, it presents classifying trending topics by considering context of trending topics. commemorative days, then it goes to commemoratives category. Otherwise, it goes to news category again.

There are some examples of trending topics classified into different categories as figure 7.5 shows. When a trending topic presents as Be + Noun, or Verb + Noun, or Person pronoun + Noun, or Possessive adjectives + Noun, it belongs to meme category. Another context pattern is noun only, which need further to divided into has commemorative days or has not. The two subordinate patterns of trending topics lead two category as table 16 shows.

As early as 2011, Zubiaga et al. introduced a typology to categorize trending topics into news, current events, memes, and commemoratives. Similarly, in this research, the categories of context of trending topics should be identified as memes, commemoratives and news. Meme refers to trending topics that diffuse through a social network often as a mimicry. Commemorative refers to trending topics about a memorial day, such as congratulating or celebrating their birthday, and the anniversary of an event or person. News refers to trending topics about breaking news or events.

After applying this feature to trending topics data, the percentage of context categories of trending topics shows as table 7.3. Only 4% of trending topics are categorized into commemoratives category and only 9% of trending topics are categorized into meme category. However, 87% of trending topics are classified into news, which are exactly same as Kwak (2010) indicated that over 85% of trending topics are related to the breaking news headlines. The context categories of trending topics are reasonable.

Rank Feature

Rank factor that represents the popularity of a trending topic utilizes ranks provided by Twitter, which contains average rank feature and starting rank feature. Since the collected trending topics data is based on the top 10 real-time trending topics provided by Twitter, each trending topic has their own ranks while the trending topics appeared in somewhere. Due to trending topics can appeared at same time in different countries, a trending topic may has different ranks because of different appeared countries. There are many ranks for a trending topic at different time in different countries, which represents the popularity of that trending topics at that time at that country. Although there are several ranks for each

trending topic, the ranks for each trending topic appeared at first time are the suitable ranks to consider. The reason why choose the ranks of starting countries for each trending topic is that the rank represents the people's highest popularity while a topic become a trending topic. There may several ranks for a trending topic that appeared at first time due to a trending topic can be started in multiple countries at the same time. The average rank feature is calculating numeric average rank of trending topics that are started in multiple countries and classifying the numeric average rank for each trending topic into different levels. The different diffusing levels of starting countries play different roles in diffusion of trending topics, which affect the ranks of trending topics. So the starting rank feature is considering the diffusing level of starting country for each trending topic.

The rank of a trending topics from started points show the popularity of initiation of this trending topic, which means the degree of people's interestingness while it becomes a trending topic. Since some of trending topics start in multiple countries, there may be several ranks for a trending topic. To be fair, the average ranking of trending topics that are started in multiple countries should be applied.

Due to the data collected based on the top 10 real-time trending topics from Twitter, the range of rank for trending topics is from rank 1 to rank 10, which means that averaging ranks of trending topics started in multiple countries should be in range of 1 to 10. Once the numeric average rank of each trending topic is calculated, it might have decimal fraction, which cannot be classified well. In order to avoid to classify the numeric average ranks directly, they need to be grouped. The numeric average ranks can be divided into three levels. The numeric average rank in range of 1 to 3 can be identified as high level of interestingness of people. The numeric average rank in range of 4 to 6 can be identified as medium level of interestingness of people. And the numeric average rank in range of 7 to 10 can be identified as low level of interestingness of people. In hence, average rank level of trending topics can be categorized into high level, medium level and low level as table 19 shows.

Table 7.4 Categories and description of average rank level

Categories	Description
High level	The numeric average rank of trending topics that appeared first time is in range of rank 1 to 3
Medium level	The numeric average rank of trending topics that appeared first time is in range of rank 4 to 6
Low level	The numeric average rank of trending topics that appeared first time is in range of rank 7 to 10

According to table 7.4, if the numeric average rank of this trending topic is in range of rank 1 to 3, then it is categorized in to high level. If the numeric average rank of this trending

topics is in range of rank 4 to 6, then it is categorized into medium level. If the numeric average rank of this trending topic is in range of rank 7 to 10, then it is categorized into low level.

Despite there are three levels of average rank of trending topics, the numeric ranks also can be considered into feature. Although average rank level of trending topics considered, which seems like fair, starting countries have their own diffusing levels, which means considering average rank level of trending topics is not enough. The different diffusing levels of starting countries for a trending topic represent the different possibility of ranks of these countries. In order to consider diffusing level of starting country, the starting rank feature should be applied. If diffusing level of starting country for a trending topic is innovator, which means the higher possibility that the trending topic started from the innovator diffused to adopters. The ranks of trending topics from innovators are more possible than the rank of trending topics of adopters. In hence, the rank of trending topics of innovator should be applied while there are several starting countries of a trending topic. In spite of the diffusing level of starting countries, there is another situation that a trending topic has different ranks, which started from the same diffusing level. Once the trending topic is started from same diffusing level of starting countries with different ranks, the highest rank should be applied. The reason why choose the highest rank because it is more appropriate to represent the popularity of this trending topic while it becomes a trending topic. In hence, based on the criteria to apply the numeric rank from original data into new one. Once a trending topic comes in, at first check whether it started in single country. If it starts in single country, then use original rank of this trending topic. If it starts in multiple countries, then check whether it starts from innovator. If it starts from innovator, then check whether there are more than one innovator for this trending topic. If more than one, then choose the highest rank of this trending topic as the starting rank. If there is only one innovator, then choose rank from innovator as starting rank of this trending topic. If there is no innovators, then choose the highest rank of this trending topic as the starting rank.

StartedCountry	Keyword	Date	Rank
ph	#10days	2013-09-28 17:15:02	2
uk	#10days	2013-09-28 17:15:02	2
usa	#10days	2013-09-28 17:15:02	1

Fig. 7.6 An example of applying starting rank feature

In order to consider the diffusing level of starting country for each trending topic, it is more complex than applying average rank level into it. Figure 7.6 shows an example of applying starting rank feature to trending topic. As we can see, the ranks of #10days are from three starting countries, rank 2 for PH and UK, rank 1 for USA. Since both diffusing levels of USA and UK are innovators, #10days is started from USA and UK possibly. The rank of #10days from PH whose diffusing level is adopted, which can be ignored. And then comparing the ranks from USA and UK. The rank from USA is rank 1 which is higher than rank 2 that is from UK. The highest rank represents higher interests when #10days are started from USA. The starting rank feature is to considering the highest diffusing level of starting country and highest rank into it, so the starting rank for #10days is rank 1. After applying the starting ranking feature into trending topics data, the percentage for rank 1 to 10. The percentage of rank 1 to 10 almost around 10% which means the trending topics for each rank occupy similar proportion in the diffusion of trending topics, which demonstrates categorization of trending topics reasonable. To model the prediction diffusion, it requires extracting prediction features and building a training data. The prediction diffusion is built with three factors which are extracted by analysing actual trending topics data. In this chapter, it will show how the system operates, and how the prediction diffusion is generated.

7.4 Evaluation Set-up

The aim of this research is to predict scale and range of diffusion of trending topics. As the figure 7.7 shows, a trending topic started in country A, and it diffused to country B, country C, country D, country E and country F. The scale of diffusion of this trending topic is six countries which includes country A to E. And the range of diffusion of this trending topic is three levels, which are that level 1 is country A, level 2 is country B and country C, and level 3 is country D, country E and country F.

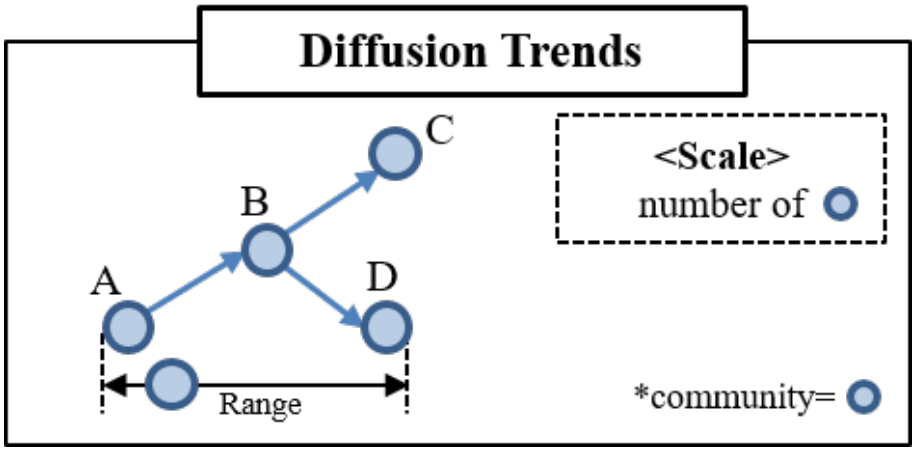

Fig. 7.7 Scale and Range of diffusion prediction

In order to predict the scale and range of trending topics, the prediction diffusion is modelled with country factor, context factor and ranking factor. The country factor is to identify the character of the country that starts each trending topics. The context factor is to classify trending topics into three different categories based on the context patterns. Then ranking factor is to classify the average rank level and calculate the starting rank for each trending topic. To evaluate the proposed prediction diffusion, five machine learning techniques are used. The first one is Naive Bayes, which is a group of simple classifiers with strong independence assumptions between the features. The second one is Neural Networks, which are computational models working like interconnected neurons. The third one is support vector machines, which use related learning algorithms to classify the boundary of classes. The fourth one is ripple down rules, which a way of approaching knowledge acquisition. The fifth one is C4.5 algorithm, which is a decision tree that referred to as a statistical classifier. We used the exactly same factors and same machine learning techniques for both two experiments. Experiment 1 and experiment 2 was performed to examine the scale prediction accuracy and range prediction accuracy for the prediction diffusion. The following experiments and results are presented including specific what we found.

7.4.1 Dataset

Trending topics data for this research are collected from eight English-speaking countries that contain the United States, the United Kingdom, Canada, Australia, New Zealand, Philippines, Malaysia, and Singapore. Each data contains country, keyword, date, rank, which are provided by Twitter. Since trending topics data is collected by every 15 minutes, the date for each data is collected at the same timing point. There are two dataset which are used as training data and testing data. The training data was collected from 8th August 2013 to 07th

November 2013, which includes 3975 unique trending topics keywords in 376077 tweets that containing these trending topics. The testing data was collected from 8th November 2013 to 7th February 2014, which includes 4286 unique trending topics keywords. The dataset was used for two experiments as follows. In the experiments, some numbers were rounded and some data was sorted to improve understanding.

7.4.2 Scale Prediction Accuracy

The main objective of this experiment is to examine the scale prediction accuracy in five learning techniques. The experiment 1 demonstrated the proposed prediction diffusion reasonably predicts the scale of diffusion of trending topics. Table 7.5 presents the scale prediction accuracy in five machine learning techniques. Using only context factor has lowest accuracy in five machine learning techniques. In contrast, using only country factor or only ranking factor has better accuracy both over 0.5. The accuracy results increased at least 0.04 up to 0.07 by using country factor and ranking factor together. All of accuracy by using three factors of evaluation in five machine learning techniques are over 0.7, which is much higher than using a single factor or two factor to predict scale.

Table 7.5 The prediction accuracy with five machine learning techniques

	Context	Rank	Country	Rank + Country	Context + Rank + Country
NB	0.212	0.582	0.545	0.632	0.722
NN	0.252	0.592	0.559	0.636	0.738
SVM	0.258	0.593	0.558	0.642	0.727
RDR	0.262	0.595	0.587	0.655	0.743
C4.5	0.261	0.599	0.588	0.666	0.748

As we can see in table 22, the accuracy results of using only ranking factor and using only country factor in NN and SVM are almost same. However, the accuracy result of using ranking factor and country factor in NN is lower than in SVM, and the accuracy result of using ranking factor, country factor and context factor in NN is higher than in SVM. SVM can solve the problem of structure selection in NN, which can improve the accuracy result of using ranking factor and country. Comparing in NN, the accuracy result of using context factor is better in SVM, the accuracy result of using three factors is lower in SVM. That means patterns for adding context factor are more fit the NN structure. Another interesting finding is that RDR and C4.5 has similar accuracy that is better than other three machine learning techniques. Since RDR and C4.5 using rule tree to classify trending topics, the classifying results may be more fit the actual trending topics data.

Since C4.5 has the most accuracy for almost factors, the scale prediction accuracy evaluating in C4.5 is analysed as an example in figure 7.8. As we can see, when we just use only context factor, the accuracy result is lower than others, which just reach 0.261. However, using only country factor or ranking factor, the accuracy result almost reach 0.6. When combining ranking factor and country factor, the accuracy result at least increased 0.07 than using single factor. The highest accuracy result almost reach 0.75 by using three factors together.

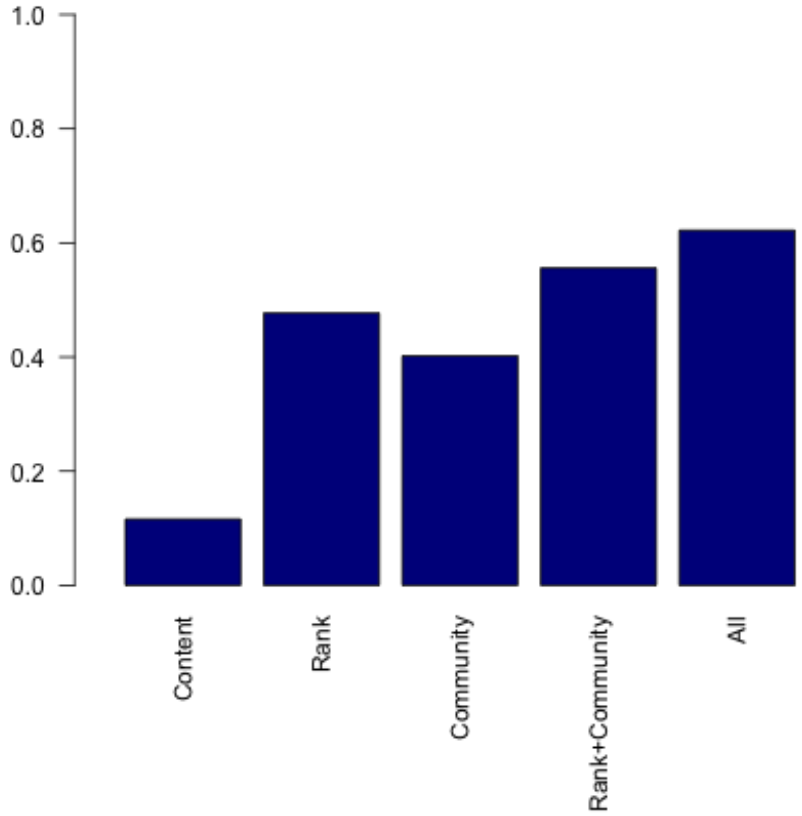

Fig. 7.8 Scale Prediction Accuracy

Using only context factor to predict, the scale prediction accuracy just reach 0.26, which is not good to predict the scale of trending topics. Since the result of analyzing trending topics data shows 87% of trending topics belongs to news category, most context categories of trending topics are same. The context factor cannot classify the results appropriately. Comparing context factor, country factor and ranking factor performed much better. Using

country factor to classify results, starting country of trending topics is classified by applying diffusing level feature and language feature. In a sense, he characters of starting country determine the diffusion route of trending topics, which makes the accuracy results high. The proportion of trending topics in each categories that applying average rank feature and starting rank feature are similar, which means the classification result of trending topics are more appropriate and prediction by using ranking factor is more reliable. Adding context factor into country factor and ranking factor almost increases 0.1 accuracy, which means context factor doesn't perform well but help rank factor and country factor perform well. In conclusion, the accuracy result of using three factor together almost reach 0.75 which is much better than predicting in randomly without factors at 0.158 accuracy, which demonstrates the prediction diffusion perform well in predicting scale of diffusion of trending topics.

7.4.3 Range Prediction Accuracy

The main objective of this experiment is to examine the scale prediction accuracy. The five machine learning techniques are applied to evaluate the accuracy for predicting range diffusion trend. Our experiment 2 demonstrated the proposed prediction diffusion reasonably predicts the range of trending topics. As we can see in table 23, NN and C4.5 has better partial accuracy than other three machine learning techniques. Using only context factor has lowest accuracy in evaluation of five machine learning techniques, which is almost 0.1. In contrast, using only country factor or ranking factor has much better accuracy than using only context factor. The accuracy result of combining rank factor and country factor together increased at least 0.1 compared to using only ranking factor or country factor in NN. It also increased at least 0.08 compared to using only ranking factor or country factor in C4.5. The accuracy result of using three factors together almost 0.1 compared to using ranking factor and country factor in NN. All accuracy results of using three factors in five machine learning techniques are over 0.6, which is much higher than using a single factor to predict, especially using only context factor

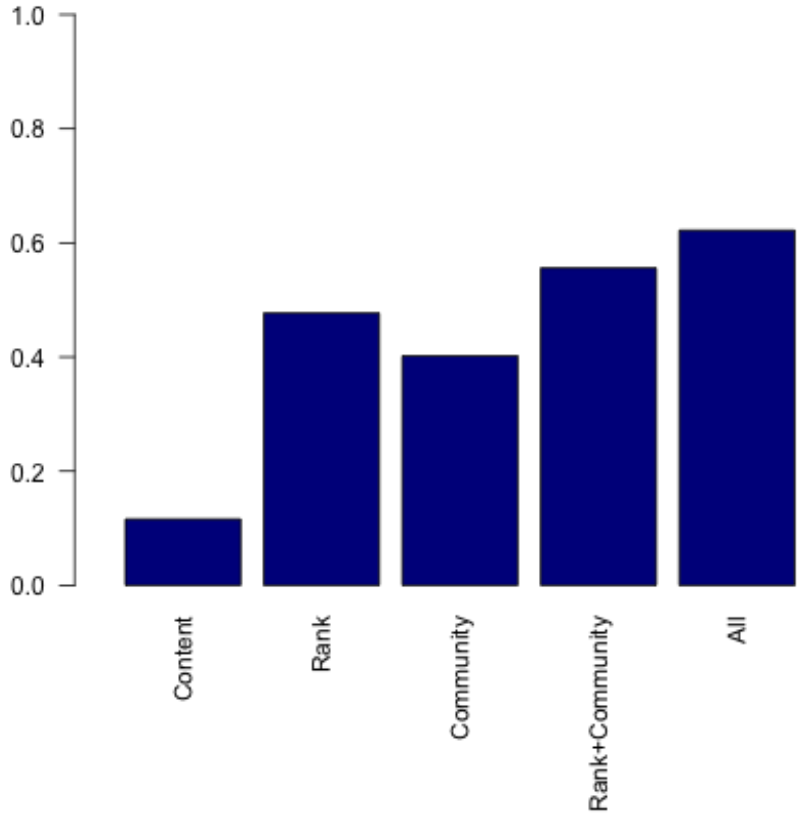

Fig. 7.9 Range Prediction Accuracy

The accuracy result of using factors performed almost better in C4.5 than in other machine learning techniques. Figure 7.9 presents the accuracy results of all factors in C4.5. As we can see, if we used only context factor, the accuracy result is lower than others, which just reach 0.12. Comparing to context factor, country factor or ranking factor has better accuracy that almost reach 0.4 to 0.5. The accuracy result increased to 0.56 when combining the ranking factor and country factor together. It reach the highest accuracy 0.622 when using three factors together. Adding context factor into using ranking factor and country factor improves almost 0.07 accuracy.

7.5 Implementation

The following figure 7.10 shows the architecture of the research.

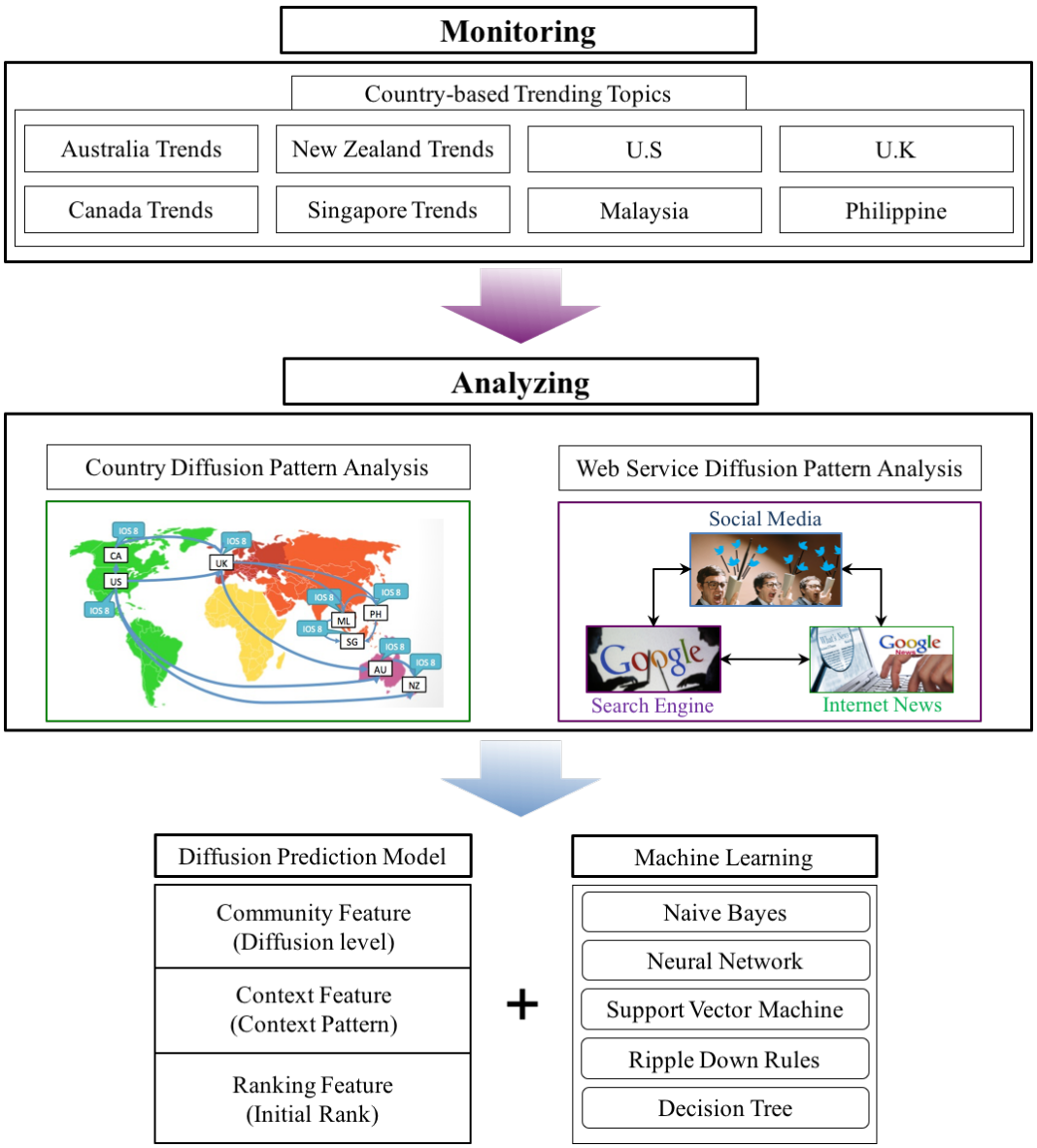

Fig. 7.10 System Architecture of predition diffusion accross countries

The database for this research is designed for storing all detailed information of trending topics from Twitter trending topics analytics services with the related information. Meaning of each stored trending topic is disambiguated by using related information and store in the database. It also contains the training data for the country diffusion prediction and accuracy results. The database for this project is designed as follows:

- Table: tb_twt_keyword
  - id: primary key, the identification number generated by auto increment function.

  - keyword: twitter trending keyword
  - rank: rank of the twitter trending keyword
  - group: group of collect time, each group has 10 keywords (1-10 Rank)
  - country: country of the twitter trending keyword
  - local_time: local time at collection
  - date: collect time

- Table: tb_twt_relatedTweets
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - tweet_id: the identification number of the related tweet given from twitter api
  - tweet_content: contents of the related tweet
  - tweet_date: uploaded time of the related tweet
  - retweet_count: the number of re-tweet of the related tweet given from twitter api
  - favorite_count: the number of favorite of the related tweet given from twitter api
  - date: collect time
  - tb_twt_relatedTweet_user_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweet_user table

- Table: tb_twt_relatedTweet_user
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_relatedTweet_id: foreign key, the identification number that enables to connect with tb_twt_relatedTweets table
  - user_id: the identification number of the twitter user given from twitter api
  - user_name: the identification username of the twitter user
  - user_screenName: the screenname of the twitter user
  - user_location: the location of the twitter user uploaded tweet
  - user_followers_count: the number of followers of the twitter user given from twitter api

  - user_friends_count: the number of friends of the twitter user given from twitter api
  - date: collect time

- Table: tb_twt_relatedNews
  - id: primary key, the identification number generated by auto increment function.
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - news_content: contents of the related news
  - news_date: uploaded time of the related news
  - source: the source of the collected related news
  - date: collect time

- Table: tb_keyword_meaning_disambiguation
  - tb_twt_keyword_id: foreign key, the identification number that enables to connect with tb_twt_keyword table
  - keyword: twitter trending keyword
  - content_tweet_kfe: key factor extraction result of the related tweets
  - content_tweet_ner: named entity reconiser result of the related tweets
  - content_tweet_tm: topic modeling result of the related tweets
  - content_tweet_as: automatic summarisation result of the related tweets
  - content_news_kfe: key factor extraction result of the related news
  - content_news_ner: named entity reconiser result of the related news
  - content_news_tm: topic modeling result of the related news
  - content_news_as: automatic summarisation result of the related news
  - content_combined_kfe: key factor extraction result of the related news and tweets
  - content_combined_ner: named entity reconiser result of the related news and tweets
  - content_combined_tm: topic modeling result of the related news and tweets
  - content_combined_as: automatic summarisation result of the related news and tweets

  - date: collect time

- Table: twt_raw_data
  - Country: country of twitter trending keyword appeared
  - Keyword: twitter trending keyword
  - Date: collect time
  - Rank: rank of the twitter trending keyword
  - Group: group of collect time, each group has 10 keywords (1-10 Rank)
  - Topic: topic of the twitter trending keyword
  - LocalTime: local time at collection
  - RelK: a set of related keywords of the twitter trending keyword

- Table: twt_training_data
  - Keyword: twitter trending keyword
  - Diffusion level: diffusion level of the twitter trending keyword
  - Language: language type of the twitter trending keyword
  - Context: context of the twitter trending keyword
  - Average Rank: average rank of the twitter trending keyword
  - Starting Rank: starting rank of the twitter trending keyword
  - Starting Scale: diffusion scale of the twitter trending keyword
  - Starting Range: diffusion range of the twitter trending keyword

- Table: twt_diffused_route
  - ID: primary key, the identification number generated by auto increment function.
  - Keyword: twitter trending keyword
  - Number: total number of diffusion
  - Level: total level of diffusion
  - Route: whole route of diffusion

7.6 Conclusion

Recently, online social services attracted people's attention, which provide topics what most people are interested in. The trending topics that provided by these online social services provide convenience for researchers to investigate and predict the people's online interestingness. Although the previous work researches the information diffusion online, there is no research about the diffusion trends of trending topics. In hence, I built a prediction diffusion to predict the diffusion trends of trending topics which include scale and range. The scale diffusion trend of trending topics presents how many countries a trending topic will diffuse. And the range diffusion trend of trending topics presents how far a diffusion chain of a trending topic can continue on in depth. This research aims to model a prediction diffusion to predict the diffusion trends of trending topics.

In order to model this prediction diffusion, I analysed three month actual trending topics data from Twitter and found three factors to identify a trending topic. They are country factor, context factor and ranking factor. The country factor is to find and categorize the characters of countries that started trending topics. There are two features to identify the characters of starting country for each trending topic. Diffusing level feature is find the role of the starting country for each trending topic in diffusion of trending topics. And language feature is to classify the starting country of trending topics by their speaking language. The context factor is to find context patterns of trending topics and categorize them into different classes. And the ranking factor represents the popularity of the trending topics, which classifies trending topics by average ranking and ranking that calculate by utilizing the ranks of each trending topic provided by Twitter.

To evaluate this model, the five machine learning techniques are used including Naive Bayes, neural network, support vector machine, ripple down rule and C4.5 algorithm. The accuracy results in these five machine learning techniques, C4.5 algorithm get the highest accuracy for predicting scale, and C4.5 and neural network get the highest accuracy for predicting range. Since C4.5 classifying trending topics by using rule tree. Once the training data was built perfectly, the prediction diffusion for predicting scale modelled by C4.5 performed perfectly. Using C4.5 as an example, the accuracy result for predicting range diffusion trend of trending topics reaches 0.62, which performed much better rather than predicting in randomly without the factors, which just reaches 0.158. The accuracy result for predicting scale diffusion trend of trending topics is better than range prediction accuracy, which reaches almost 0.75. The reason that predicting scale gets the better accuracy is that the prediction factors are more suitable for predicting scale than predicting range. Besides, our assumption for this research is that we consider just once for each country that appear a trending topic. In hence, there may chance that a trending topic started in USA and diffused to

UK and CA, and then went back to US. However, we just consider as a trending topic started in USA and diffused to UK and CA. Therefore, the range diffusion trend that we consider may differentiate with the actual range diffusion trend, which may leads the lower accuracy for prediction. In conclusion, our experiments were performed successfully, especially for predicting scale of the diffusion of trending topics.

Chapter 8

Study Conclusion

This research investigated the nature of trending topics and proposed the related smart service framework using trending topics. The main objectives were to identify possibilities and limitations inherent in the nature of current trending topics analytics service and to provide novel recommendation and prediction frameworks that can be used for individuals or organisation in different research or industry fields.

8.1 Summary of Contributions

Chapter 2 explored the possibilities and overview of prediction and detection research using web social data and trending topics. The previous prediction research works using web social data were covered various aims, including election prediction, disaster detection, and disease prediction/detection. Those researches represent the possibilities of web social data. The chapter also introduced several types of trending topic analytics services, which display the most popular searched, discussed or read topics by the users in web services, such as search engine, social media, or internet news site. The data from trending topic analytics service is the main resource for the thesis.

Chapter 3 is to characterise the trending topics by tracking the trending topics in different countries and services. Chapter 3 aimed to identify the most successful method to retrieve the representative contents for twitter trending topics sense disambiguation. In order to achieve this, four different information retrieval approaches are evaluated, including key factor extraction, named entity recognition, topic modelling, and automatic summarisation, by human experiments with 20 postgraduate students. The results in the chapter shows that statistical key factor extraction approach, a classical term weighting technique, provides the highest performance in retrieving the most representative contents for trending topics sense disambiguation. The research present the result of the first human evaluation in online

trending topic sense disambiguation. The best meaning disambiguation approach, term frequency, is used in the chapter 3,4,5,6, and 7. This approach revealed the idea of the specific trending topic in order to identify the relevance and predict the future trends of a trending topic.

Chapter 4 is to identify the relevance of trending topic to a target object, such as individuals or organisations. Chapter 4 proposed a smart service framework that identifies the relevance between trending topics and a target object, such as an individual or organisation. The result proved in the evaluation that the system can extract the accurate related keyword from Twitter and Internet news. The advantage of extracting four related keywords is shown. In order to find the relevance, it is crucial to construct the virtual target domain that is well-structured and contains up to date information. Despite remaining some uncertainties about how the method should be implemented, it can be seen that the proposed system, personalised relevance identification system, is a valuable service.

Chapter 5 to 7 are to develop the model that predicts the trends of trending topics in the future. Chapter 5 addressed trending topic rank prediction problems. The research suggests a simple rank prediction that uses historical data with consideration of window size and missing value treatment. Surprisingly, the method achieved very significant performance (about 94% accuracy with C4.5 decision tree). On the one hand, this implicitly implies that the changing trends are the most important factors for rank prediction. On the other hand, it would be possible to improve performance of rank prediction. However, it would be very difficult to predict rank perfectly (100% accuracy), which is not because of algorithmic factors but because of trending topics' irregularly changing nature. The research is the initial work that proposes the temporal model of predicting of trending topics ranking as a degree of people's interests, and it achieves the successful result.

Online trending topics show popular trending topics in certain online community. The communities can be countries or web services. Based on the analysis result, we found that trending topics in one country are different from others, and some of them diffuse through multiple communities. Chapter 7 and 8 proposed new framework for predicting the diffusion trends of trending topics among different online communities.

Chapter 6 proposed a model that predicted the diffusion trends of trending topics among different web services. The research also examined how the events/issues diffuse among those three web services, including search engine, social media, and internet news. The research provides a characterization of trending topic diffusion based on topic, time, and service that can be used as features for prediction model. The performance was evaluated by applying 8 types of machine learning techniques, including Fuzzy Unordered Rule Induction Algorithm (FURIA), Support Vector Machine (SVM), Knearest neighbour (KNN), C4.5

Decision Tree (C4.5), Ripple Down Rules (RDR), Kstar, Feed Forward Neural Network (FFNN), Logistic Regression (LR). The research is the first ever study on the trending topic diffusion model through the web services. Trending topic diffusion model has never been reported before in the literatures.

Chapter 7 is the initial research of modelling to predict the diffusion trends, including scale and range, of trending topics among different countries. The research found that over 90% of trending topic for each country are appeared in different countries. For example, 92.27% of trending topics in UK are appeared in at least one other countries. (only 7.73% of trending topic in UK appeared only in UK) It represents that the trending topics are shared in not only one but various countries.It predicted how many countries that a trending topic diffuses and how far the diffusion chain of a trending topic can continue. The evaluation result found the diffusion prediction model with provider factor, context factor and ranking factor achieved high performance.

8.2 Recommendation for Future Research

This dissertation focused on investigating the dynamics of trending topics, and proposing new type of smart services framework using trending topics. Since the trending topic represents the people's interests in the real-time, it would be worth to adopt this framework to the real-world social problems, such as stock forecasting, marketing and election prediction. The trending topic lifecycle and diffusion prediction model proposed in this thesis can be one of the valuable attributes for predicting the future activity in some domains, which are affected by people or social issue.

8.2.1 Stock Prediction with Trending Topics

Stock market prediction is the area that focuses on determining the future value/price of a company stock or other financial instruments traded on an exchange. The stock price is reflected by all newly revealed information or social issues so it is inherently unpredictable. In order to develop a successful prediction model of a stock's future price, it would be useful to identify the factor that represents the relationship between stock price and revealed social issues. For the future work, we will build a new framework by combining the proposed trending topics lifecycle prediction model and relevance identification model. This combined framework enables to identify the relevance of trending topics to a company, which provides a chance for stakeholders to decide their actions in the stock market. Moreover, the trending topic lifecycle prediction model in this framework can estimate the degree of change of

popularity in the trending topic, and reveal the future influence of the trending topics to the company stock.

Appendix A

Database Schema

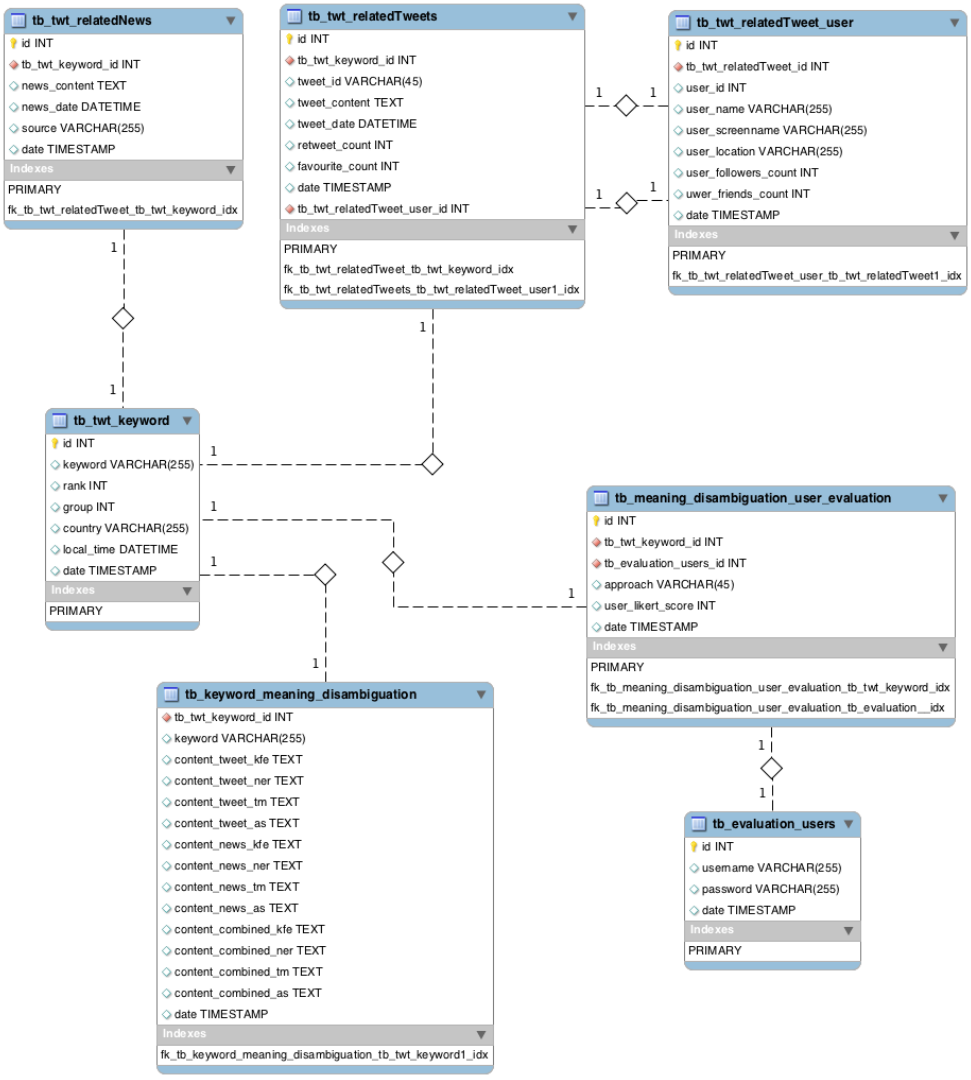

Fig. A.1 Database Schema for Chapter 3

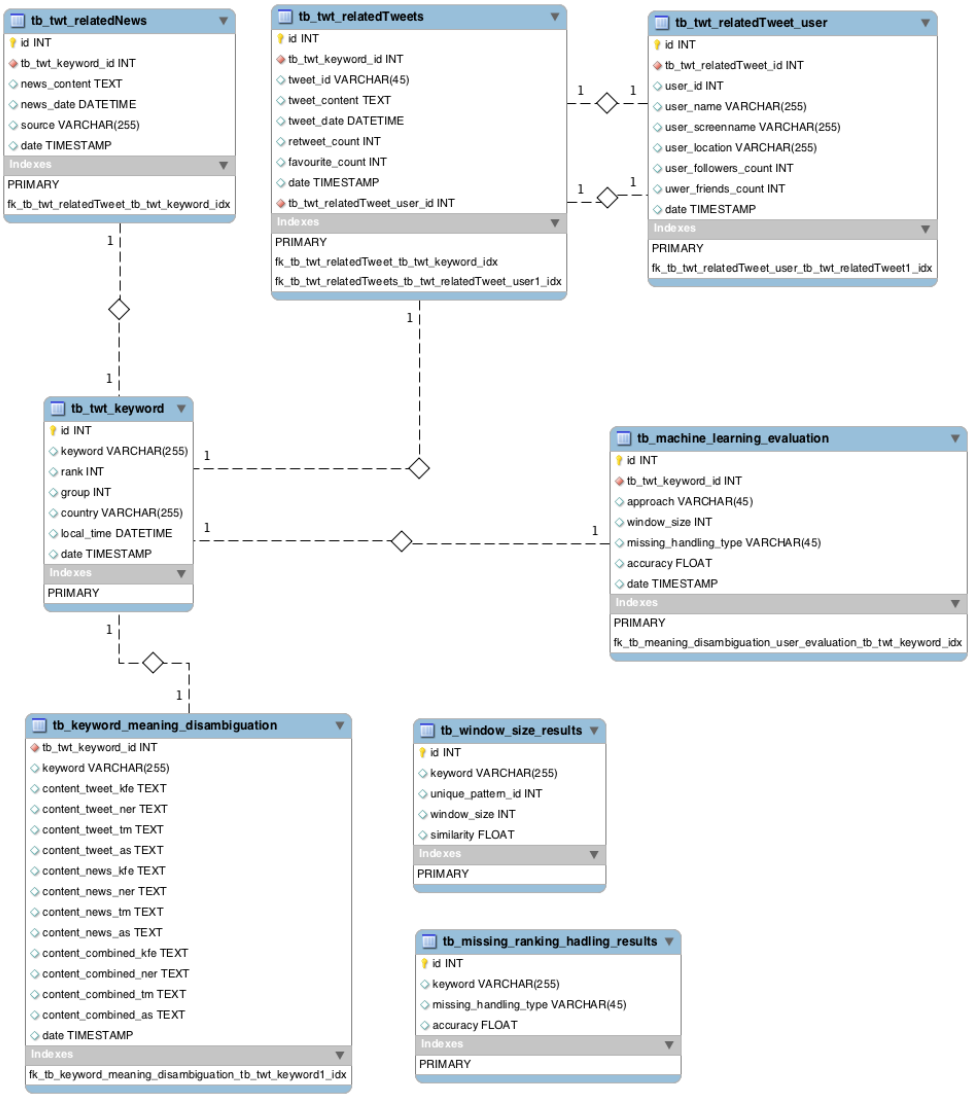

Fig. A.2 Database Schema for Chapter 5

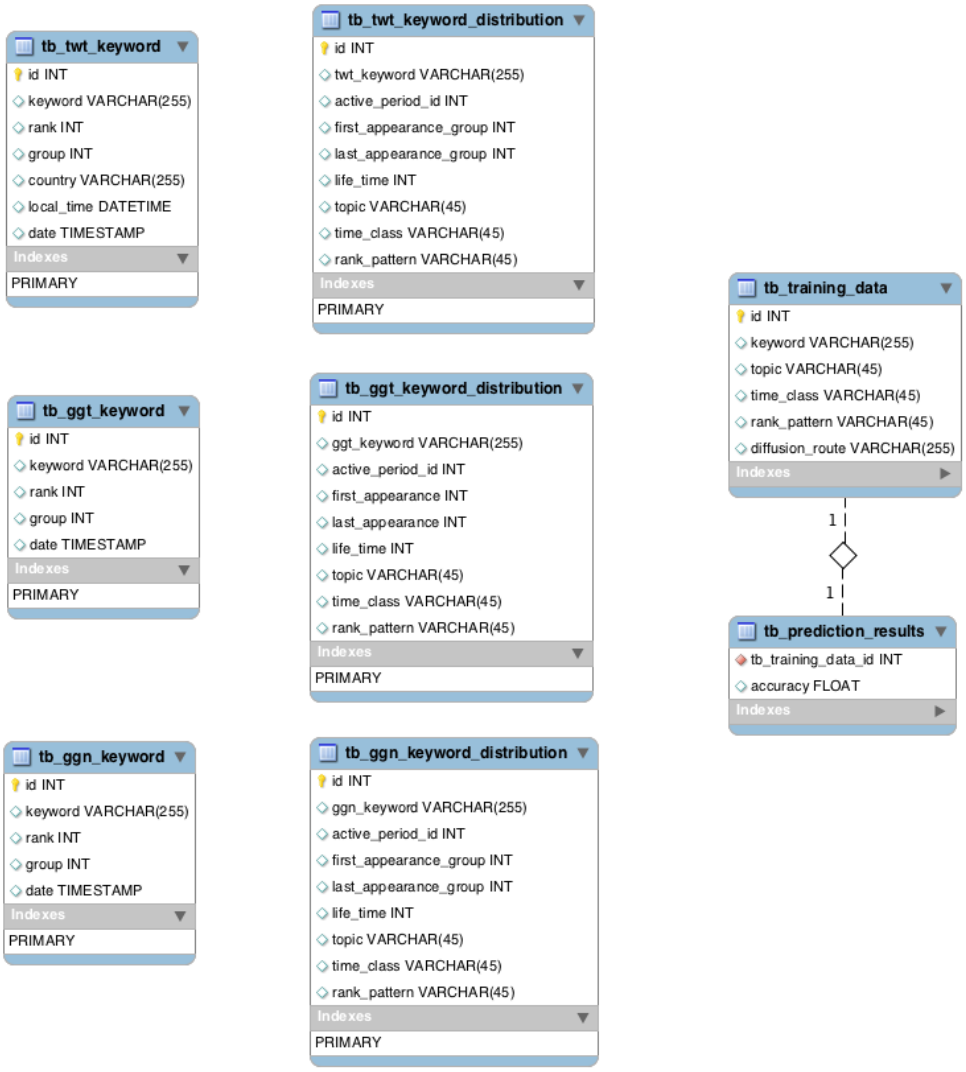

Fig. A.3 Database Schema for Chapter 6

Appendix B

Australia Government website List for Website Monitoring

URL	Number of Documents
http://www.daff.gov.au/	1960
http://www.communications.gov.au/	546
http://www.defence.gov.au/	281
http://www.ag.gov.au/	289
http://www.education.gov.au/	517
http://www.employment.gov.au/	137
http://www.finance.gov.au/	145
http://www.dfat.gov.au/	659
http://www.health.gov.au/	382
http://www.humanservices.gov.au/	485
http://www.immi.gov.au/	273
http://www.industry.gov.au/	861
http://www.infrastructure.gov.au/	274
http://www.dss.gov.au/	242
http://www.environment.gov.au/	252
http://www.dpmc.gov.au/	185
http://www.dva.gov.au/	990
http://www.treasury.gov.au/	885

Fig. B.1 Australia Government website List for website monitoring

Appendix C

Twitter Trending Topics Daily Log- Sample

Date	Twitter Trending Topic Collection Daily Log
13/11/2012	960
14/11/2012	960
15/11/2012	960
16/11/2012	960
17/11/2012	960
18/11/2012	960
19/11/2012	960
20/11/2012	960
21/11/2012	960
22/11/2012	960
23/11/2012	960
24/11/2012	960
25/11/2012	960
26/11/2012	960
27/11/2012	960
28/11/2012	960
29/11/2012	960
30/11/2012	960
1/12/2012	960
2/12/2012	960
3/12/2012	960
4/12/2012	960
5/12/2012	960
6/12/2012	960
7/12/2012	960
8/12/2012	960
9/12/2012	960
10/12/2012	960
11/12/2012	960
12/12/2012	960
13/12/2012	960
14/12/2012	960
15/12/2012	960
16/12/2012	960
17/12/2012	960
18/12/2012	960
19/12/2012	960

Fig. C.1 Twitter Trending Topics Daily Log - Sample

Bibliography

Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011, June). Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In Proceedings of the 3rd International Web Science Conference (p. 2). ACM.

Achrekar, H., A. Gandhe, R. Lazarus, S.-H. Yu and B. Liu (2011). Predicting flu trends using twitter data. Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on, IEEE.

Aiello, L. M., G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris and A. Jaimes (2013).“Sensing trending topics in Twitter.”

Al Bawab, Z., G. H. Mills and J.-F. Crespo (2012). Finding trending local topics in search queries for personalization of a recommendation system. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

AL-Mutairi, H. M. and M. B. Khan (2014). Predicting the Popularity of Trending Articles in the Arabic Wikipedia using Data Mining Techniques. Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems, ACM.

Althoff, T., D. Borth, J. Hees and A. Dengel (2013). Analysis and forecasting of trending topics in online media streams. Proceedings of the 21st ACM international conference on Multimedia, ACM.

Aly, A. A. (2008).“Using a query expansion technique to improve document retrieval.” International Journal “Information Technologies and Knowledge 2(4): 343-348.

Andersen, P. M., Hayes, P. J., Huettner, A. K., Schmandt, L. M., Nirenburg, I. B., & Weinstein, S. P. (1992, March). Automatic extraction of facts from press releases to generate news stories. In Proceedings of the third conference on Applied natural language processing (pp. 170-177). Association for Computational Linguistics.

Aramaki, E., S. Maskawa and M. Morita (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Archambault, D., D. Greene and P. Cunningham (2013).“TwitterCrowds: Techniques for Exploring Topic and Sentiment in Microblogging Data.”arXiv preprint arXiv:1306.3839.

Asur, S. and B. A. Huberman (2010).“Predicting the future with social media.”arXiv preprint arXiv:1003.5699.

Asur, S., B. A. Huberman, G. Szabo and C. Wang (2011). Trends in social media: persistence and decay. ICWSM.

Bakshy, E., I. Rosenn, C. Marlow and L. Adamic (2012). The role of social networks in information diffusion. Proceedings of the 21st international conference on World Wide Web, ACM.

Barbosa, L. and J. Feng (2010). Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics.

Bar-Haim, R., E. Dinur, R. Feldman, M. Fresko and G. Goldstein (2011). Identifying and following expert investors in stock microblogs. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Becker, H., M. Naaman and L. Gravano (2011).“Beyond Trending Topics: Real-World Event Identification on Twitter.”ICWSM 11: 438-441.

Benhardus, J. and J. Kalita (2013).“Streaming trend detection in twitter.”International Journal of Web Based Communities 9(1): 122-139.

Bermingham, A. and A. F. Smeaton (2010). Classifying sentiment in microblogs: is brevity an advantage? Proceedings of the 19th ACM international conference on Information and knowledge management, ACM.

Bermingham, A. and A. F. Smeaton (2011).“On using Twitter to monitor political sentiment and predict election results.”

Bian, J., Y. Yang and T.-S. Chua (2013). Multimedia summarization for trending topics in microblogs. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, ACM.

Bifet, A. and E. Frank (2010). Sentiment knowledge discovery in twitter streaming data. Discovery Science, Springer.

Bollen, J., H. Mao and X. Zeng (2011).“Twitter mood predicts the stock market.”Journal of Computational Science 2(1): 1-8.

Bollen, J., A. Pepe and H. Mao (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Brennan, M. and R. Greenstadt (2011). Coalescing twitter trends: The under-utilization of machine learning in social media. Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom), IEEE.

Breyer, B. N., Sen, S., Aaronson, D. S., Stoller, M. L., Erickson, B. A., & Eisenberg, M. L. (2011). Use of Google Insights for Search to track seasonal and geographic kidney stone incidence in the United States. Urology, 78(2), 267-271.

Cao, R., Liang, X., & Ni, Z. (2012). Stock Price Forecasting with Support Vector Machines Based on Web Financial Information Sentiment Analysis. In Advanced Data Mining and Applications (pp. 527-538). Springer Berlin Heidelberg.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clinical infectious diseases, 49(10), 1557-1564.

Carrascosa, J. M., González, R., Cuevas, R., & Azcorra, A. (2013, October). Are trending topics useful for marketing?: visibility of trending topics vs traditional advertisement. In Proceedings of the first ACM conference on Online social networks (pp. 165-176). ACM.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010, July). Emerging topic detection on twitter based on temporal and social terms evaluation. In Proceedings of the Tenth International Workshop on Multimedia Data Mining (p. 4). ACM.

Centola, D. (2010). The spread of behavior in an online social network experiment. science, 329(5996), 1194-1197.

Chang, H. C. (2010). A new perspective on Twitter hashtag use: Diffusion of innovation theory. Proceedings of the American Society for Information Science and Technology, 47(1), 1-4.

Chen, L., Zhang, C., & Wilson, C. (2013, October). Tweeting under pressure: analyzing trending topics and evolving word choice on sina weibo. In Proceedings of the first ACM conference on Online social networks (pp. 89-100). ACM.

Cheong, M. (2009). 'What are you Tweeting about?': A survey of Trending Topics within Twitter. Clayton School of Information Technology, Monash University.

Cheong, M., & Lee, V. (2009, November). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In Proceedings of the 2nd ACM workshop on Social web search and mining (pp. 1-8). ACM.

Choi, D., Hwang, M., Kim, J., Ko, B., & Kim, P. (2014). Tracing trending topics by analyzing the sentiment status of tweets. Computer Science and Information Systems, 11(1), 157-169.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), 2-9.

Choy, M., Cheong, M. L., Laik, M. N., & Shung, K. P. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520.

Chum, O., Mikulik, A., Perdoch, M., & Matas, J. (2011, June). Total recall II: Query expansion revisited. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 889-896). IEEE.

Chung, J. E., & Mustafaraj, E. (2011, August). Can collective sentiment expressed on twitter predict political elections?. In AAAI (Vol. 11, pp. 1770-1771).

Chung, S., & Liu, S. (2011). Predicting stock market fluctuations from twitter. Berkeley, California.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003, August). A comparison of string metrics for matching names and records. In Kdd workshop on data cleaning and object consolidation (Vol. 3, pp. 73-78).

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011, October). Predicting the political alignment of twitter users. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 192-199). IEEE.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. ICWSM, 133, 89-96.

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. International journal of environmental research and public health, 7(2), 596-615.

Cui, A., Zhang, M., Liu, Y., Ma, S., & Zhang, K. (2012, October). Discover breaking events with popular hashtags in twitter. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1794-1798). ACM.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. arXiv preprint arXiv:1007.4748.

Culotta, A. (2010, July). Towards detecting influenza epidemics by analyzing Twitter messages. In Proceedings of the first workshop on social media analytics (pp. 115-122). ACM.

De Choudhury, M., Lin, Y. R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media?. ICWSM, 10, 34-41.

De Choudhury, M., Lin, Y. R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media?. ICWSM, 10, 34-41.

Diakopoulos, N. A., & Shamma, D. A. (2010, April). Characterizing debate performance via aggregated twitter sentiment. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1195-1198). ACM.

Diaz-Aviles, E., Drumond, L., Gantner, Z., Schmidt-Thieme, L., & Nejdl, W. (2012, October). What is happening right now... that interests me?: online topic discovery and recommendation in twitter. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1592-1596). ACM.

Dow, P. A., Adamic, L. A., & Friggeri, A. (2013, June). The Anatomy of Large Facebook Cascades. In ICWSM.

Earle, P. (2010). Earthquake twitter. Nature Geoscience, 3(4), 221-222.

Earle, P. S., Bowden, D. C., & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics, 54(6).

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer Mediated Communication, 13(1), 210 230.

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2013, October). Traveling trends: social butterflies or frequent fliers?. In Proceedings of the first ACM conference on Online social networks (pp. 213-222). ACM.

Fiaidhi, J., Mohammed, S., & Islam, A. (2012). Towards identifying personalized twitter trending topics using the twitter client rss feeds. Journal of Emerging Technologies in Web Intelligence, 4(3), 221-226.

Fiaidhi, J., Mohammed, S., Islam, A., Fong, S., Kim, T. H., Sharma, S., & Bhushan, B. (2013). Developing a hierarchical multi-label classifier for Twitter trending topics. International Journal of u-and e-Service, Science and Technology, 6(3), 1-12.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 363-370). Association for Computational Linguistics.

Fitzpatrick, K. (2007). The pleasure of the blog: The early novel, the Serial, and the narrative archive.

Foster, S., Potter, W., Wu, J., Hu, B., & Zhang, Y. (2009, March). A history sensitive cascade model in diffusion networks. In Proceedings of the 2009 Spring Simulation Multiconference (p. 5). Society for Computer Simulation International.

Gao, D., Li, W., Cai, X., Zhang, R., & Ouyang, Y. (2014). Sequential summarization: A full view of twitter trending topics. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 22(2), 293-302.

Gilbert, E., & Karahalios, K. (2009, April). Predicting tie strength with social media. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 211-220). ACM.

Giummolè, F., Orlando, S., & Tolomei, G. (2013, September). Trending topics on Twitter improve the prediction of Google hot queries. In Social Computing (SocialCom), 2013 International Conference on (pp. 39-44). IEEE.

Goetz, M., Leskovec, J., McGlohon, M., & Faloutsos, C. (2009, May). Modeling Blog Dynamics. In ICWSM.

Granovetter, M. (1978). Threshold models of collective behavior. American journal of sociology, 1420-1443.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ quality & safety, bmjqs-2012.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010, October). @ spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security (pp. 27-37). ACM.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004, May). Information diffusion through blogspace. In Proceedings of the 13th international conference on World Wide Web (pp. 491-501). ACM.

Guha, R., Kumar, R., Raghavan, P., & Tomkins, A. (2004, May). Propagation of trust and distrust. In Proceedings of the 13th international conference on World Wide Web (pp. 403-412). ACM.

Gupta, A., & Kumaraguru, P. (2012, April). Credibility ranking of tweets during high impact events. In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (p. 2). ACM.

Gupta, A., & Kumaraguru, P. (2012, April). Credibility ranking of tweets during high impact events. In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media (p. 2). ACM.

Gutierrez, C. E., Alsharif, M. R., & Yamashita, K. (2014). Uncover Trending Topics on Data Stream by Linear Prediction Modeling. International Journal of Computer Science and Network Security (IJCSNS), 14(5), 1.

Han, S. C., & Chung, H. (2012). Social issue gives you an opportunity: Discovering the personalised relevance of social issues. In Knowledge management and acquisition for intelligent systems (pp. 272-284). Springer Berlin Heidelberg.

Han, S. C., Chung, H., & Kang, B. H. (2012). It is time to prepare for the future: forecasting social trends. In Computer applications for database, education, and ubiquitous computing (pp. 325-331). Springer Berlin Heidelberg.

Han, S. C., Chung, H., Kim, D. H., Lee, S., & Kang, B. H. (2014). Twitter trending topics meaning disambiguation. In Knowledge management and acquisition for smart systems and services (pp. 126-137). Springer International Publishing.

Han, S. C., & Kang, B. H. (2012, June). Identifying the relevance of Social Issues to a Target. In Web Services (ICWS), 2012 IEEE 19th International Conference on (pp. 666-667). IEEE.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (2012, April). Discovering geographical topics in the twitter stream. In Proceedings of the 21st international conference on World Wide Web (pp. 769-778). ACM.

Hong, L., Dom, B., Gurumurthy, S., & Tsioutsiouliklis, K. (2011, August). A time-dependent topic model for multiple text streams. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 832-840). ACM.

Hossmann, T., Carta, P., Schatzmann, D., Legendre, F., Gunningberg, P., & Rohner, C. (2011, December). Twitter in disaster mode: security architecture. In Proceedings of the Special Workshop on Internet and Disasters (p. 7). ACM.

Hossmann, T., Legendre, F., Carta, P., Gunningberg, P., & Rohner, C. (2011, September). Twitter in disaster mode: Opportunistic communication and distribution of sensor data in emergencies. In Proceedings of the 3rd Extreme Conference on Communication: The Amazon Expedition (p. 1). ACM.

Hsieh, C. C., Moghbel, C., Fang, J., & Cho, J. (2013). Experts vs the crowd: Examining popular news prediction perfomance on twitter. WWW. ACM.

Inouye, D., & Kalita, J. K. (2011, October). Comparing twitter summarization algorithms for multiple post summaries. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 298-306). IEEE.

Irani, D., Webb, S., Pu, C., & Li, K. (2010). Study of trend-stuffing on twitter through text classification. In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS).

Iyengar, A., Finin, T., & Joshi, A. (2011, October). Content-based prediction of temporal boundaries for events in Twitter. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 186-191). IEEE.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 151-160). Association for Computational Linguistics.

Jin, X., Gallagher, A., Cao, L., Luo, J., & Han, J. (2010, October). The wisdom of social multimedia: using flickr for prediction and forecast. In Proceedings of the international conference on Multimedia (pp. 1235-1244). ACM.

Joinson, A. N. (2008, April). Looking at, looking up or keeping up with people?: motives and use of facebook. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 1027-1036). ACM.

Soler, J. M., Cuartero, F., & Roblizo, M. (2012, August). Twitter as a tool for predicting elections results. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (pp. 1194-1200). IEEE Computer Society.

Juang, Y. S., Lin, S. S., & Kao, H. P. (2008). A knowledge management system for series-parallel availability optimization and design. Expert Systems with Applications, 34(1), 181-193.

Kairam, S. R., Morris, M. R., Teevan, J., Liebling, D. J., & Dumais, S. T. (2013, June). Towards Supporting Search over Trending Events with Social Media. In ICWSM.

Kang, B. H., Kim, D. H., & Chung, H. (2014, January). What issue spread on the web: analyze the web trends. In Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication (p. 24). ACM.

Kang, M., Zhong, H., He, J., Rutherford, S., & Yang, F. (2013). Using google trends for influenza surveillance in South China. PloS one, 8(1), e55205.

Kempe, D., Kleinberg, J., & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 137-146). ACM.

Kim, H., Beznosov, K., & Yoneki, E. (2014, April). Finding influential neighbors to maximize information diffusion in twitter. In Proceedings of the companion publication of the 23rd international conference on World Wide Web companion (pp. 701-706). International World Wide Web Conferences Steering Committee.

Kim, M., Newth, D., & Christen, P. (2013, August). Modeling direct and indirect influence across heterogeneous social networks. In Proceedings of the 7th Workshop on Social Network Mining and Analysis (p. 9). ACM.

Kinsella, S., Murdock, V., & O'Hare, N. (2011, October). I'm eating a sandwich in Glasgow: modeling locations with tweets. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 61-68). ACM.

Kireyev, K., Palen, L., & Anderson, K. (2009, December). Applications of topics models to analysis of disaster-related twitter data. In NIPS Workshop on Applications for Topic Models: Text and Beyond (Vol. 1). Canada: Whistler.

Kolbitsch, J., & Maurer, H. A. (2006). The Transformation of the Web: How Emerging Communities Shape the Information we Consume. J. UCS, 12(2), 187-213.

Kossinets, G., Kleinberg, J., & Watts, D. (2008, August). The structure of information pathways in a social communication network. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 435-443). ACM.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11, 538-541.

Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011, July). TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In ICWSM.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM.

Kwon, J., & Han, I. (2013, January). Information diffusion with content crossover in online social media: An empirical analysis of the social transmission process in twitter. In System Sciences (HICSS), 2013 46th Hawaii International Conference on (pp. 3292-3301). IEEE.

Kwon, Y. S., Kim, S. W., & Park, S. (2009, November). An analysis of information diffusion in the blog world. In Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management (pp. 27-30). ACM.

Kwon, Y. S., Kim, S. W., Park, S., Lim, S. H., & Lee, J. B. (2009, June). The information diffusion model in the blog world. In Proceedings of the 3rd Workshop on Social Network Mining and Analysis (p. 4). ACM.

Laber, E. S., de Souza, C. P., Jabour, I. V., de Amorim, E. C. F., Cardoso, E. T., Rentería, R. P., ... & Valentim, C. D. (2009, November). A fast and simple method for extracting relevant content from news webpages. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1685-1688). ACM.

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society, 14(5), 729-747.

Lau, J. H., Collier, N., & Baldwin, T. (2012). On-line Trend Analysis with Topic Models:# twitter Trends Detection Topic Model Online. In COLING (pp. 1519-1534).

Lee, C., Kwak, H., Park, H., & Moon, S. (2010, April). Finding influentials based on the temporal order of information adoption in twitter. In Proceedings of the 19th international conference on World wide web (pp. 1137-1138). ACM.

Lee, C. H., Chien, T. F., & Yang, H. C. (2011, October). An automatic topic ranking approach for event detection on microblogging messages. In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on (pp. 1358-1363). IEEE.

Lee, C. H., Yao, H., He, X., Chan, S. H., Chang, J., & Maghoul, F. (2014, April). Learning to predict trending queries: classification-based. In Proceedings of the companion publication of the 23rd international conference on World wide web companion (pp. 335-336). International World Wide Web Conferences Steering Committee.

Lee, K., Caverlee, J., Kamath, K. Y., & Cheng, Z. (2012, April). Detecting collective attention spam. In Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality (pp. 48-55). ACM.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011, December). Twitter trending topic classification. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on (pp. 251-258). IEEE.

Lerman, K., & Ghosh, R. (2010). Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. ICWSM, 10, 90-97.

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010, April). Predicting positive and negative links in online social networks. In Proceedings of the 19th international conference on World wide web (pp. 641-650). ACM.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., & Hurst, M. (2007, April). Patterns of Cascading behavior in large blog graphs. In SDM (Vol. 7, pp. 551-556).

Li, D., Xu, Z., Luo, Y., Li, S., Gupta, A., Sycara, K., ... & Chen, H. (2013, October). Modeling information diffusion over social networks for temporal dynamic prediction. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 1477-1480). ACM.

Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012, April). Tedas: A twitter-based event detection and analysis system. In Data engineering (icde), 2012 ieee 28th international conference on (pp. 1273-1276). IEEE.

Li, Z., Wang, B., Li, M., & Ma, W. Y. (2005, August). A probabilistic model for retrospective news event detection. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 106-113). ACM.

Lilien, G. L., Rangaswamy, A., & Bruyn, A. D. (2007). The Bass model: marketing engineering technical note. University of Washington, Seattle.

Lim, S. H., Kim, S. W., Kim, S., & Park, S. (2011, March). Construction of a blog network based on information diffusion. In Proceedings of the 2011 ACM Symposium on Applied Computing (pp. 937-941). ACM.

Lin, Y. C., Yang, P. C., Hsieh, W. T., & Seng-cho, T. C. (2012). Technology Trend Analysis Tool using Twitter as a Source. In Proceedings of International Conference on Information Technology, E-Government and Application.

Liu, F., Liu, Y., & Weng, F. (2011, June). Why is sxsw trending?: exploring multiple text sources for twitter topic summarization. In Proceedings of the Workshop on Languages in Social Media (pp. 66-75). Association for Computational Linguistics.

Liu, J., Dolan, P., & Pedersen, E. R. (2010, February). Personalized news recommendation based on click behavior. In Proceedings of the 15th international conference on Intelligent user interfaces (pp. 31-40). ACM.

Liu, Y., Han, W., Tian, Y., Que, X., & Wang, W. (2013, November). Trending topic prediction on social network. In Broadband Network & Multimedia Technology (IC-BNMT), 2013 5th IEEE International Conference on (pp. 149-154). IEEE.

Lobzhanidze, A., Zeng, W., Gentry, P., & Taylor, A. (2013, January). Mainstream media vs. social media for trending topic prediction-an experimental study. In Consumer Communications and Networking Conference (CCNC), 2013 IEEE (pp. 729-732). IEEE.

Ma, T., & Wan, X. (2010, August). Opinion target extraction in Chinese news comments. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 782-790). Association for Computational Linguistics.

Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in twitter. Journal of the American Society for Information Science and Technology, 64(7), 1399-1410.

Mackie, S., McCreadie, R., Macdonald, C., & Ounis, I. (2014). Comparing algorithms for microblog summarisation. In Information Access Evaluation. Multilinguality, Multimodality, and Interaction (pp. 153-159). Springer International Publishing.

Malik, M. T., Gumel, A., Thompson, L. H., Strome, T., & Mahmud, S. M. (2011). " Google flu trends" and emergency department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique, 294-297.

Mao, Y., Wei, W., Wang, B., & Liu, B. (2012, August). Correlating S&P 500 stocks with Twitter data. In Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research (pp. 69-72). ACM.

Martin, C., Corney, D., & Goker, A. (2015). Mining newsworthy topics from social media. In Advances in Social Media Analysis (pp. 21-43). Springer International Publishing.

Martinez-Romo, J., & Araujo, L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40(8), 2992-3000.

Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 1155-1158). ACM.

Mccord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In Autonomic and trusted computing (pp. 175-186). Springer Berlin Heidelberg.

McGlohon, M., Leskovec, J., Faloutsos, C., Hurst, M., & Glance, N. (2007). Finding patterns in blog shapes and blog evolution.

Medvet, E., & Bartoli, A. (2012, December). Brand-related events detection, classification and summarization on twitter. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on (Vol. 1, pp. 297-302). IEEE.

Mendoza, M., Poblete, B., & Castillo, C. (2010, July). Twitter Under Crisis: Can we trust what we RT?. In Proceedings of the first workshop on social media analytics (pp. 71-79). ACM.

Mills, A., Chen, R., Lee, J., & Raghav Rao, H. (2009). Web 2.0 emergency applications: how useful can Twitter be for emergency response?. Journal of Information Privacy and Security, 5(3), 3-26.

Mukherjee, S., Sujithan, R., & Subasic, P. (2014, April). Detecting trending topics using page visitation statistics. In Proceedings of the companion publication of the 23rd international conference on World wide web companion (pp. 347-348). International World Wide Web Conferences Steering Committee.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. Journal of the American Society for Information Science and Technology, 62(5), 902-918.

Narayan, S., Prodanovic, S., Elahi, M. F., & Bogart, Z. (2010). Population and Enrichment of Event Ontology using Twitter. Information Management SPIM 2010, 31.

Nikolov, S., & Shah, D. (2012, November). A nonparametric method for early detection of trending topics. In Proceedings of the Interdisciplinary Workshop on Information and Decision in Social Networks (WIDS 2012). MIT.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. ICWSM, 11(122-129), 1-2.

Oh, C., & Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement.

Okazaki, M., & Matsuo, Y. (2010). Semantic twitter: analyzing tweets for real-time event notification. In Recent Trends and Developments in Social Software (pp. 63-74). Springer Berlin Heidelberg.

Paolillo, J. C. (2008, January). Structure and network in the YouTube core. In Hawaii International Conference on System Sciences, Proceedings of the 41st Annual (pp. 156-156). IEEE.

Park, S. S., Kim, Y. S., & Kang, B. H. (2003). Web Information Management System: Personalization and Generalization. In ICWI (p. 532).

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. ICWSM, 20, 265-272.

Pervin, N., Fang, F., Datta, A., Dutta, K., & Vandermeer, D. (2013). Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. ACM Transactions on Management Information Systems (TMIS), 3(4), 19.

Petrović, S., Osborne, M., & Lavrenko, V. (2010, June). Streaming first story detection with application to twitter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 181-189). Association for Computational Linguistics.

Phuvipadawat, S., & Murata, T. (2010, August). Breaking news detection and tracking in Twitter. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 3, pp. 120-123). IEEE.

Pohl, D., Bouchachia, A., & Hellwagner, H. (2012, April). Automatic sub-event detection in emergency management using social media. In Proceedings of the 21st international conference companion on World Wide Web (pp. 683-686). ACM.

Popescu, A. M., & Pennacchiotti, M. (2010, October). Detecting controversial events from twitter. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1873-1876). ACM.

Qiu, L., Rui, H., & Whinston, A. B. (2011). A twitter-based prediction market: Social network approach. Available at SSRN 2047846.

Rao, T., & Srivastava, S. (2014). Twitter sentiment analysis: How to hedge your bets in the stock markets. In State of the Art Applications of Social Network Analysis (pp. 227-247). Springer International Publishing.

Radas, S. (2006). Diffusion Models in Marketing: How to Incorporate the Effect of External Influence?. Privredna kretanja i ekonomska politika, 15(105), 30-51.

Rech, J. (March 2007). Discovering trends in software engineering with google trend. SIGSOFT Softw. Eng. Notes 32, 2, 1-2.

Recuero, R., & Araújo, R. (2012, June). On the rise of artificial trending topics in twitter. In Proceedings of the 23rd ACM conference on Hypertext and social media (pp. 305-306). ACM.

Reis, D. D. C., Golgher, P. B., Silva, A. S., & Laender, A. (2004, May). Automatic web news extraction using tree edit distance. In Proceedings of the 13th international conference on World Wide Web (pp. 502-511). ACM.

Remy, C., Pervin, N., Toriumi, F., & Takeda, H. (2013, December). Information Diffusion on Twitter: everyone has its chance, but all chances are not equal. In Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on (pp. 483-490). IEEE.

Rightler-McDaniels, J. L., & Hendrickson, E. M. (2014). Hoes and hashtags: constructions of gender and race in trending topics. Social Semiotics, 24(2), 175-190.

Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(5), 522-532.

Ritter, A., Clark, S., & Etzioni, O. (2011, July). Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1524-1534). Association for Computational Linguistics.

Rogers, E. M. (2010). Diffusion of innovations. Simon and Schuster.

Rogstadius, J., Kostakos, V., Laredo, J., & Vukovic, M. (2011). Towards real-time emergency response using crowd supported analysis of social media. In Proceedings of CHI workshop on crowdsourcing and human computation, systems, studies and platforms.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011, March). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In Proceedings of the 20th international conference on World wide web (pp. 695-704). ACM.

Rosa, H., Carvalho, J. P., & Batista, F. (2014). Detecting a tweet's topic within a large number of Portuguese Twitter trends. In OASIcs-OpenAccess Series in Informatics (Vol. 38). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Ruan, Y., Purohit, H., Fuhry, D., Parthasarathy, S., & Sheth, A. P. (2012). Prediction of topic volume on twitter.

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012, February). Correlating financial time series with micro-blogging activity. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 513-522). ACM.

Russell, M. A. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. O'Reilly Media, Inc..

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009, November). Twitterstand: news in tweets. In Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems (pp. 42-51). ACM.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.

Sharifi, B., Hutton, M. A., & Kalita, J. (2010, January). Automatic summarization of twitter topics. In National Workshop on Design and Analysis of Algorithm, Tezpur, India.

Sharifi, B., Hutton, M. A., & Kalita, J. (2010, June). Summarizing microblogs automatically. In Human Language Technologies: The 2010 Annual Conference of the North American

Chapter of the Association for Computational Linguistics (pp. 685-688). Association for Computational Linguistics.

Sharifi, B., Hutton, M. A., & Kalita, J. K. (2010, August). Experiments in microblog summarization. In Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 49-56). IEEE.

Soman, S. J., & Murugappan, S. (2014, July). Bayesian probabilistic tensor factorization for malicious tweets in trending topics. In Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on (pp. 895-900). IEEE.

Song, X., Tseng, B. L., Lin, C. Y., & Sun, M. T. (2006, August). Personalized recommendation driven by information flow. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 509-516). ACM.

Sprenger, T. O. (2011, July). Tweettrader. net: Leveraging crowd wisdom in a stock microblogging forum. In Fifth International AAAI Conference on Weblogs and Social Media.

Stafford, G., & Yu, L. L. (2013, September). An Evaluation of the Effect of Spam on Twitter Trending Topics. In Social Computing (SocialCom), 2013 International Conference on (pp. 373-378). IEEE.

Story, J., & Wickstra, J. (2011). Discovering trending topics on twitter via retweets. Unpublished manuscript. Retrieved from http://cs. uiowa. edu/ jwikstr/finalPaper. pdf.

Subašić, I., & Berendt, B. (2011). Peddling or creating? investigating the role of twitter in news reporting. In Advances in Information Retrieval (pp. 207-213). Springer Berlin Heidelberg.

Sun, E., Rosenn, I., Marlow, C., & Lento, T. M. (2009, May). Gesundheit! Modeling Contagion through Facebook News Feed. In ICWSM.

Taxidou, I., & Fischer, P. (2013). Realtime analysis of information diffusion in social media. Proceedings of the VLDB Endowment, 6(12), 1416-1421.

Teevan, J., Ramage, D., & Morris, M. R. (2011, February). #TwitterSearch: a comparison of microblog search and web search. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 35-44). ACM.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Tirado, J. M., Higuero, D., Isaila, F., & Carretero, J. (2011, April). Analyzing the impact of events in an online music community. In Proceedings of the 4th Workshop on Social Network Systems (p. 6). ACM.

Tran, T., Georgescu, M., Zhu, X., & Kanhabua, N. (2014, June). Analysing the duration of trending topics in Twitter using Wikipedia. In Proceedings of the 2014 ACM conference on Web science (pp. 251-252). ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10, 178-185.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. Social Science Computer Review, 0894439310386557.

Villena Román, J., Lana Serrano, S., Martínez Cámara, E., & González Cristóbal, J. C. (2013). TASS-Workshop on sentiment analysis at SEPLN.

Wang, A. H. (2010, July). Don't follow me: Spam detection in twitter. In Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on (pp. 1-10). IEEE.

Wang, S., Paul, M. J., & Dredze, M. (2014). Exploring health topics in Chinese social media: An analysis of Sina Weibo. AAAI Work World Wide Web Public Heal Intell.

Wang, W., & Wu, B. (2011, June). Comparing Twitter and Chinese native microblog. In Cybersecurity Summit (WCS), 2011 Second Worldwide (pp. 1-4). IEEE.

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In Social Computing, Behavioral-Cultural Modeling and Prediction (pp. 231-238). Springer Berlin Heidelberg.

Weinshall, D., Levi, G., & Hanukaev, D. (2013). Lda topic model with soft assignment of descriptors to words. In Proceedings of the 30th International Conference on Machine Learning (pp. 711-719).

Weng, J., & Lee, B. S. (2011). Event Detection in Twitter. ICWSM, 11, 401-408.

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. Journal of the American Society for Information Science and Technology, 63(8), 1631-1646.

Wong, F. M. F., Sen, S., & Chiang, M. (2012, August). Why watching movie tweets won't tell the whole story?. In Proceedings of the 2012 ACM workshop on Workshop on online social networks (pp. 61-66). ACM.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3), 13.

Wu, S., Gong, L., Rand, W., & Raschid, L. (2012, September). Making recommendations in a microblog to improve the impact of a focal user. In Proceedings of the sixth ACM conference on Recommender systems (pp. 265-268). ACM.

Wukich, C., & Steinberg, A. (2013). Nonprofit and Public Sector Participation in Self Organizing Information Networks: Twitter Hashtag and Trending Topic Use During Disasters. Risk, Hazards & Crisis in Public Policy, 4(2), 83-109.

Xia, Y., Yu, H., & Zhang, S. (2009, November). Automatic web data extraction using tree alignment. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1645-1648). ACM.

Xu, W., Ritter, A., Callison-Burch, C., Dolan, W. B., & Ji, Y. (2014). Extracting lexically divergent paraphrases from Twitter. Transactions of the Association for Computational Linguistics, 2, 435-448.

Yang, C. C., Jiang, L., Yang, H., & Tang, X. (2012, August). Detecting signals of adverse drug reactions from health consumer contributed content in social media. In Proceedings of ACM SIGKDD Workshop on Health Informatics.

Yang, C. C., Yang, H., Jiang, L., & Zhang, M. (2012, October). Social media mining for drug safety signal detection. In Proceedings of the 2012 international workshop on smart health and wellbeing (pp. 33-40). ACM.

Yang, J., & Counts, S. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. ICWSM, 10, 355-358.

Yang, J., & Leskovec, J. (2010, December). Modeling information diffusion in implicit networks. In Data Mining (ICDM), 2010 IEEE 10th International Conference on (pp. 599-608). IEEE.

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 27(6), 52-59.

Yu, L., Asur, S., & Huberman, B. A. (2011). What trends in Chinese social media. arXiv preprint arXiv:1107.3522.

Yu, L., Asur, S., & Huberman, B. A. (2012). Artificial Inflation: The True Story of Trends in Sina Weibo. arXiv preprint. arXiv, 1202.

Yu, S., & Kak, S. (2012). A survey of prediction using social media. arXiv preprint arXiv:1203.1647.

Zhang, C. M., & Paxson, V. (2011, March). Detecting and analyzing automated activity on twitter. In Passive and Active Measurement (pp. 102-111). Springer Berlin Heidelberg.

Zhang, R., Li, W., Gao, D., & Ouyang, Y. (2013). Automatic twitter topic summarization with speech acts. Audio, Speech, and Language Processing, IEEE Transactions on, 21(3), 649-658.

ZHANG, W., ZHENG, N., REN, Y., XU, J., ZHANG, H., & XU, M. (2014). Discovery of Trending Topics in Microblog Streams Based on Contextual Search. Journal of Computational Information Systems, 10(2), 491-498.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, 26, 55-62.

Zhang, Y., & Pennacchiotti, M. (2013, May). Predicting purchase behaviors from social media. In Proceedings of the 22nd international conference on World Wide Web (pp. 1521-1532). International World Wide Web Conferences Steering Committee.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In Advances in Information Retrieval (pp. 338-349). Springer Berlin Heidelberg.

Zhou, E., Haoran, L., Sun, H., & Sun, L. (2013). U.S. Patent Application No. 13/904,445.

Zubiaga, A., Spina, D., Fresno, V., & Martínez, R. (2011, October). Classifying trending topics: a typology of conversation triggers on twitter. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2461-2464). ACM.